Consultative Committee for Space Data Systems

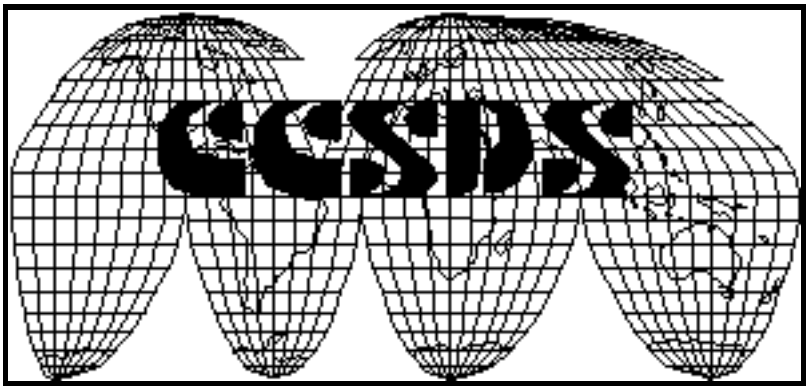
RECOMMENDATION CONCERNING SPACE DATA SYSTEMS STANDARDS

Reference Model for an Open Archival Information System (OAIS)

CCSDS 650.0-W-2.0

WHITE BOOK

October 15, 1997



Dear Reader:

The following version of the OAIS Reference Model is designated CCSDS 650.0-W-2.0. This version, called White Book 2, reflects most of the updates, both specific and general, agreed to at the ISO Silver Spring International Workshop held in May, 1997.

Comments may be sent to:

Lou Reich
louis.i.reich@gsfc.nasa.gov

or

Don Sawyer
donald.sawyer@gsfc.nasa.gov

The next version of this book is anticipated for release in January, 1998 and a final version is planned for May, 1998.

AUTHORITY

Issue:	White Book, Issue 2.0
Date:	October 15, 1997
Location:	Silver Spring, USA

This document, when it has been approved for publication by the Management Council of the Consultative Committee for Space Data Systems (CCSDS), will reflect the consensus of technical panel experts from CCSDS Member Agencies. The procedure for review and authorization of CCSDS Reports is detailed in reference [1].

This document is published and maintained by:

CCSDS Secretariat
Program Integration Division (Code MG)
National Aeronautics and Space Administration
Washington, DC 20546, USA

FOREWORD

This document is a technical Recommendation for use in developing a consensus on what is required for an archive to provide permanent, or indefinite long-term, preservation of digital information. It may be useful as a starting point for a similar document addressing the indefinite long-term preservation of non-digital information.

This Recommendation establishes a common framework of terms and concepts which comprise an Open Archival Information System (OAIS). It allows existing and future archives to be more meaningfully compared and contrasted. It provides a basis for further standardization within an archival context and it should promote greater vendor awareness of, and support of, archival requirements.

Through the process of normal evolution, it is expected that expansion, deletion, or modification to this document may occur. This Recommendation is therefore subject to CCSDS document management and change control procedures, which are defined in Reference [1].

DOCUMENT CONTROL

Document	Title	Date	Status and Substantive Changes
CCSDS 650.0-W-1	Report Concerning Space Data Systems Standards: Reference Model for an Open Archival Information System (OAIS)	April 1997	Original Issue
CCSDS 650.0-W-1.1		July 1997	Revised information model and reduced text and complexity in Section 2. Partial response to Silver Spring Workshop directions.
CCSDS 650.0-W-1.2		Sept. 1997	Further revision of information model to incorporate concept of packaging information,. More complete response to Silver Spring Workshop directions.
CCSDS 650.0-W-2.0		Oct. 1997	Complete revisions from Silver Spring Workshop instructions.

CONTENTS

Section **Page**

1 INTRODUCTION 1

1.1 PURPOSE AND SCOPE 1

1.2 APPLICABILITY 2

1.3 RATIONALE 2

1.4 ROAD-MAP FOR DEVELOPMENT OF RELATED STANDARDS 3

1.5 DOCUMENT STRUCTURE 3

1.6 DEFINITIONS 5

1.6.1 ACRONYMS AND ABBREVIATIONS 5

1.6.2 TERMS 6

1.7 REFERENCES 11

2 OAIS CONCEPTS 13

2.1 OAIS ENVIRONMENT 14

2.2 OAIS INFORMATION 14

2.2.1 INFORMATION DEFINITION 14

2.2.2 INFORMATION PACKAGE DEFINITION 16

2.2.3 INFORMATION PACKAGE VARIANTS 17

2.3 OAIS HIGH LEVEL EXTERNAL INTERACTIONS 18

2.3.1 PRODUCER INTERACTION 19

2.3.2 CONSUMER INTERACTION 19

2.3.3 MANAGEMENT INTERACTION 20

3 OAIS RESPONSIBILITIES 21

3.1 NEGOTIATES AND ACCEPTS INFORMATION 21

3.2 DETERMINES DESIGNATED CONSUMER COMMUNITIES 22

3.3 ENSURES INFORMATION IS INDEPENDENTLY USABLE 23

3.4 ASSUMES SUFFICIENT CONTROL FOR PRESERVATION 24

3.5 FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES 25

3.6 MAKES THE INFORMATION AVAILABLE 26

4 DETAILED MODELS 28

4.1 FUNCTIONAL MODEL 28
4.1.1 COMMON SERVICES 30
4.1.2 INGEST 30
4.1.3 ARCHIVAL STORAGE 31
4.1.4 DATA MANAGEMENT 32
4.1.5 ADMINISTRATION 33
4.1.6 ACCESS 34
4.1.7 DISSEMINATION 34
4.1.8 DATA FLOW AND CONTEXT DIAGRAMS 35

4.2 INFORMATION MODEL 38
4.2.1. BACKGROUND CONCEPTS 38
4.2.2. LOGICAL MODEL FOR ARCHIVAL INFORMATION 41
4.2.3 LOGICAL MODEL OF INFORMATION IN AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS) 47

4.3 HIGH LEVEL DATA FLOWS AND TRANSFORMATIONS 57
4.3.1 DATA TRANSFORMATIONS IN THE PRODUCER ENTITY 59
4.3.2. DATA TRANSFORMATIONS IN THE INGEST FUNCTIONAL AREA 59
4.3.3. DATA TRANSFORMATIONS BY THE STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS 60
4.3.4. DATA FLOWS AND TRANSFORMATIONS IN THE ACCESS FUNCTIONAL AREA 60
4.3.5. DATA FLOWS AND TRANSFORMATIONS IN THE DISSEMINATION PROCESS 61

5 MIGRATION PERSPECTIVES 63

5.1 REPLICATION 64

5.2 REPACKAGING 64

5.3 TRANSMUTATION 65

5.4 UPDATING THE PRESERVATION DESCRIPTION INFORMATION 66

6 ARCHIVE CLASSIFICATIONS 67

7 ILLUSTRATIVE SCENARIOS 70

ANNEX A. SCENARIOS OF EXISTING ARCHIVES 74

A.1 PLANETARY DATA SYSTEM ARCHIVE 74

A.2 NATIONAL ARCHIVES AND RECORDS ADMINISTRATION'S CENTER FOR ELECTRONIC RECORDS 78

A.3 LIFE SCIENCES DATA ARCHIVE 85

A.4 NATIONAL COLLABORATIVE PERINATEL PROJECT (NCPP) 1959-1974 90

A.5 ARCHIVE SCENARIO FOR THE *CENTRE DES DONNEES DE LA PHYSIQUE DES PLASMAS* (CDPP) 95

ANNEX B. FEDERATION OF ARCHIVES 102

ANNEX C. ENTITY AND FUNCTION MATRIX 106

ANNEX D. COMPATIBILITY WITH OTHER STANDARDS 107

ANNEX E. BRIEF GUIDE TO THE OMT 108

1 INTRODUCTION

1.1 PURPOSE AND SCOPE

The purpose of this document is to define the ISO Reference Model for an **Open Archival Information System** (OAIS). An OAIS is a type of **archive**, consisting of an organization of people and systems, that has accepted the responsibility to preserve information for one or more designated communities.

In this reference model there is a particular focus on digital information, both as the primary forms of information held and as supporting information for both digitally and physically archived materials. Therefore the model accommmodates information that is inherently non-digital (e.g. a physical sample) but the modeling and preservation of such information is not addressed in detail. The information being maintained has been deemed to need **Long-term Preservation**, even if the OAIS itself is not permanent. **Long Term** is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, and it extends indefinitely. This reference model:

- provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access.

- provides the concepts needed by non-archival organization to be effective participants in the preservation process.

- provides a framework that facilitates the description and comparison of the architectures and operations of existing and future archives.

- provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time.

- provides a foundation that may be expanded by other efforts to cover long-term preservation of information that is NOT in digital form (e.g., physical media, physical samples).

- expands consensus on the elements and processes for long-term digital information preservation and access, and it promotes a larger market which vendors can support.

- guides the identification and production of OAIS related standards.

The ISO Reference Model for an Open Archival Information System defines a minimum set of responsibilities for the recognition of an OAIS archive. This allows an OAIS archive to be distinguished from other uses of the term 'archive.' Various classification criteria for an OAIS are provided to allow distinguishing types of archives which may have significantly different management requirements and/or significantly different services.

The Reference Model addresses a full range of archival information preservation functions including ingest, archival storage, access, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It identifies both internal and external interfaces to the archive functions, and it identifies a number of high level services at these interfaces.

1.2 APPLICABILITY

The OAIS model in this document is applicable to organizations with the responsibility of making information available for the long term. This model is also of interest to those organizations and individuals who create information that may need long-term preservation and those that may need to acquire information from such archives.

This model may also be useful for those organizations developing or operating shorter-term archives, or repositories, for two reasons:

– When taking into consideration the rapid pace of technology, there is the likelihood that many repositories, thought of as temporary, will in fact find that some or much of their holdings will need the same type of attention as that given by permanent archives.

– Although some repositories may be temporary, some or all of their information may need to be preserved indefinitely. Such repositories need to be active participants in the long-term preservation effort.

Standards developers are expected to use this model as a basis for further standardization and therefore provide an extension of what is meant by "operating an OAIS archive". A large number of related standards are possible. A road-map for such development is briefly addressed in Section 1.4.

This Reference Model does not specify an implementation. Actual implementations may group or break out functionality differently.

1.3 RATIONALE

A tremendous growth in computational power, and in networking bandwidth and connectivity, have resulted in an explosion of organizations who are making information available in electronic forms. Transactions among all types of organizations are being conducted using electronic forms that are taking the place of more traditional forms such as paper.

Preserving information in electronic forms is much more difficult than for forms such as paper and film. This is not only a problem for traditional archives, but for many organizations that have never thought of themselves as performing an archival function. It is expected that this reference model, by establishing minimum requirements for an OAIS archive along with a set of archival concepts, will provide a common framework from which

to view archival challenges, particularly as they relate to digital information. This should enable more organizations to understand the issues and to take proper steps to ensure long term information preservation. It should also provide a basis for more standardization and therefore a larger market that vendors can support in meeting archival requirements.

1.4 ROAD-MAP FOR DEVELOPMENT OF RELATED STANDARDS

This Reference Model serves to identify areas suitable for the development of OAIS related standards. Some of these standards may be developed by Panel 2 of the CCSDS (sub-committee of ISO); others may be developed by other standardization bodies. However, any such work undertaken by other bodies should be coordinated with CCSDS Panel 2 in order to minimize incompatibilities and efforts. Areas for potential OAIS related standards include:

- standard(s) for the interfaces between OAIS type archives.
- standard(s) for the submission (ingest) of digital data sources to the archive.
- standard(s) for the delivery of digital sources from the archive.
- standard(s) for the submission of digital metadata, about digital or physical data sources, to the archive. Here, it can be envisaged that different disciplines might require different metadata standards.
- standard(s) for the identification of digital data within the archive
- protocol standard(s) to search and retrieve metadata information about digital and physical data sources.
- standard(s) for media access allowing replacement of media management systems without having to re-write the media
- standard(s) for specific physical media
- standard(s) for the migration of information across media and representations
- standard(s) for recommended archival practices and accreditation of archives

1.5 DOCUMENT STRUCTURE

Section 2 provides a high level overview of the major concepts involved in an OAIS archive. It provides a view of the environment of an OAIS archive and the roles played by those who interact with it. It discusses what is meant by “information” and what is necessary to preserve it for the long term.

Section 3 defines mandatory responsibilities an OAIS archive must discharge in preserving its information.

Section 4 provides model views needed for a detailed understanding of an OAIS archive. It breaks down the OAIS into a number of functional areas and it identifies some high level services at the interfaces It also provides detailed data model views of information using OMT diagrams.

Section 5 provides some perspectives on the issue or migration of information across media

and across new formats or representations.

Section 6 provides some characteristics by which archives may be categorized.

Section 7 provides scenarios, using the terms and concepts defined previously, to show how information may flow into and out of an archive, and how information may be migrated within an archive.

Annex A provides scenarios of existing archive operations.

Annex B is an introduction to OAIS federations.

Annex C maps the relationships of functions within the data flow model of the OAIS.

Annex D relates parts of this reference model to other standards work.

Annex E provides a brief tutorial on the Object Modeling Technique (OMT).

1.6 DEFINITIONS

1.6.1 ACRONYMS AND ABBREVIATIONS

AIC - Archival Information Collection
AIP - Archival Information Package
AIU - Archival Information Unit
ASCII - American Standard Code for Information Interchange
CAD - Computer-Automated Design
CCSDS - Consultative Committee for Space Data Systems
CD-ROM - Compact Disk - Read Only Memory
CI - Content Information
CIP - Catalog Inter-operability Protocol
CRC - Cyclical Redundancy Check
CT - Computer Tomograpy
DED - Data Entity Dictionary
DBMS - Data Base Management System
DDL - Data Description Language
DIP - Dissemination Information Package
DR - Descriptive Record
DVD - Digital Video Disk
EBCDIC - Extended Binary Coded Decimal Interchange Code
FITS - Flexible Image Transfer System
GIF - Graphics Interchange Format
HFMS - Hierarchical File Management System
IEEE - Institute of Electrical and Electronic Engineers
IO - Information Object
IP - Information Package
ISBN - International Standard Book Number
ISO - Organization for International Standardization
LSDA - Life Sciences Data Archive
NARA - National Archives and Records Administration
NASA - National Aeronautics and Space Administration
NSSDC - National Space Science Data Center
OAIS - Open Archival Information System
ODL - Object Description Language
OMT - Object Modeling Technique
PDI - Preservation Description Information
PDMP - Project Data Management Plan
PDS - Planetary Data System
PSDD - Planetary Science Data Dictionary
PI - Packaging Information
SIP - Submission Information Package
UNICODE - Universal Code
WWW - World-Wide Web

1.6.2 TERMS

Access: This OAIS entity contains the services and functions which make the archival information and externally-available services visible to Consumers.

Access Aids: Software or documents that allow Consumers to locate, analyze, and order Archival Information Packages of interest.

Access Methods: A method for retrieving an Archival Information Package based on its name or identifier which is available to authorized users.

Active Archive: An archive where data is flowing regularly into the archive over an extended period of time at a rate driven by the Producer. This data is rapidly made accessible to Consumers.

Adhoc Request: A request that is generated by a Consumer for information the OAIS has indicated is currently available.

Administration: This OAIS entity contains the services and functions needed to control the operation of the other OAIS functional entities on a moment by moment basis

Archive: A repository that intends to preserve information for access and use by one or more Designated Communities.

Archival Storage: This OAIS entity contains the services and functions used for the storage and retrieval of Archival Information Packages.

Archival Information Collection (AIC): An Archival Information Package whose Content Information is an aggregation of other Archival Information Packages.

Archival Information Package (AIP): An information packaging concept that requires the presence of Content Information and all the associated Preserving Description Information that is needed to preserve the Content Information over the long term. It has associated Packaging Information.

Archival Information Unit (AIU): An Archival Information Package whose Content Information is not further broken down into other Content Information components, each of which has its own complete Preservation Description Information. It can be viewed as an "atomic" AIP. An example of an AIU would be a table of numbers representing temperatures in a certain region with all the associated documentation describing how and where the temperatures were measured, what instruments were used to make the measurements, who made the measurements, why they were made, what processing has been performed on the measurements and who has had custody of these measurements since they were first created, how the measurements relates to other information, how the measurement can be uniquely referenced by others, etc.

Associated Descriptions: Information describing the content of an Information Package from the point of view of a particular Acess Aid.

Client: An application which exchanges information with another application.

Collection Description: An Associated Description that described the collection as a whole.

Collection Descriptor : A type of Package Descriptor that is specialized to provide searchable information about a collection.

Common Services: Supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, and directory services necessary to support the OAIS.

Consumer: The role played by those persons, or client systems, who interact with OAIS services to find preserved information of interest and to access that information in detail.

Content Information (CI): That set of information that is the primary target for preservation. It is distinguished from Preservation Description Information which is used to assist in the preservation of the Content Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures but it excludes the documentation which would explain its history and origin, how it relates to other observations, etc.

Context Information: Information that documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere.

Data: The representation forms of information. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the sounds made by a person speaking, a moon rock.

Data Delivery Session: May be a delivered set of media or a single telecommunications session. The Data Delivery Session format/contents is based on a data model negotiated between the OAIS and the producer in the Submission Agreement. This data model identifies the logical constructs used by the producer and how they are represented on each media delivery or in the telecommunication session.

Data Dissemination Session: May be a delivered set of media or a single telecommuncations session. The Data Dissemination Session format/contents is based on a data model negotiated between the OAIS and the Consumer in the Request Agreement. This data model identifies the logical constructs used by the OAIS and how they are represented on each media delivery or in the telecommunication session.

Data Dictionary: A formal repository of terms used to describe data.

Data Management: This OAIS entity contains the services and functions for populating, maintaining, and querying a wide variety of information such as catalogs and inventories on what may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data (if any), consumer access statistics, security controls, OAIS schedules and procedures.

Data Management Data: Data created and stored in Data Management persistent storage that refer to operation of an archive. Some examples of this data are accounting data for consumer billing and authorization, policy data, subscription data for repeating requests, and statistical data for generating reports to archive management.

Data Object: Either a Physical Object or a Digital Object.

Designated Community: An identification of a set of potential Consumers who should be able to understand a particular set of information.

Descriptive Information: That set of information, consisting primarily of Package Descriptors, which is provided to Data Management to support the finding of preserved information by Consumers.

Digital Object: An object composed of a set of bit sequences.

Dissemination: The act of providing the Dissemination Information Package to the Consumer.

Dissemination Information Package (DIP): An Information Package that contains parts or all of one or more AIPs, that is distributed to the Consumer as requested.

Finding Aid: A type of Access Aid that allows a user to search for and identify Archival Information Packages of interest.

Fixity Information: This information documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner. An example is the CRC code for a file.

Format: The sequential organization of data in terms of its components.

Independently Usable Information: Information with sufficient documentation to allow the data to be understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.

Information: Any type of knowledge that can be exchanged. In an exchange, it is represented by data. Often the representation used is not fully known to the recipient of the data and the data must be accompanied by explicit Reprentation Information, understandable to the recipient, that is used to interpret the data. An example is a string of bits (the data) accompanied by a description of how to interpret a string of bits as numbers representing

temperature observations measured in degrees Celsius (the representation information).

Information Object (IO): A physical or digital object together with optional Representation Information.

Information Package: An information packaging concept that distinguishes Content Information from associated Preservation Description Information where the Preservation Description Information applies to the Content Information and is needed to aid in the preservation of the Content Information. It has associated Packaging Information used to delimit and identify the Content Information and Preservation Description Information.

Ingest: This entity contains the services and functions that accept Submission Information Packages from Producers, prepare Archival Information Packages for storage, and ensure that Archival Information Packages and their supporting Catalog Information become established within the OAIS.

Long Term: A period of time which is long enough to be concerned about the impacts of changing technologies, including support for new media and data formats, on the information being held in a repository. This period extends into the indefinite future.

Long-term Preservation: The act of preserving information, in a form which can be made understandable to a Designated Community, over the Long Term.

Management: Management is the role played by those who set overall OAIS policy as one component in a broader policy domain.

Member Description: An Associated Description that describes a member of a collection.

Metadata: Data about other data.

Migration: The act of moving a digital Information Object across storage media using either Replication, Repackaging, or Transmutation.

Object Layer Information: Information that provides additional meaning beyond that provided by the Structure Layer.

Open Archival Information System (OAIS): An OAIS is a type of **archive**, consisting of an organization of people and systems, that has accepted the responsibility to preserve information for one or more designated communities. It accepts the responsiblities defined in Section 3 of this document, and it adhers to future OAIS standards as they are defined.

Packaging Information (PI): That information that is stated as being used to bind and identify the components of an Information Package. For example, it may be the ISO-9660 volume and directory information used on a CD-ROM to provide the content of several files containing Content Information and Preservation Description Information.

Physical Object: An object (such as a moon rock, biospecimen, microscope slide) with physically observable properties that represent information that is considered suitable for being adequately documented for preservation, distribution and independent usage.

Preserve: Maintain information, in a correct and Independently Usable form, over the Long Term.

Preservation Description Information (PDI): Information necessary to adequately preserve the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information.

Producer. The role played by those persons, or client systems, who provide the information to be preserved.

Provenance Information: Information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of Provenance Information are the principal investigator who recorded the data and the information concerning its storage, handling and migration.

Reference Information: Information that identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides these identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. An example of reference information is an ISBN.

Reference Model: A framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

Repackaging: When moving a digital Information Object to new media, re-organize the Information Object on the media without requiring any alteration to its Representation Information. For example, consider the case of an Information Object whose data are the bit sequences within two files on a CD-ROM written under ISO-9660. and where its Representation Information documents the meaning of these two bit sequences. These two bit sequences may be copied to digital linear tape as two new files as long as the association of these files content to the corresponding Representation Information is maintained.

Replication: When moving a digital Information Object to new media, do a bit-for-bit copy so that the new media is virtually indistinguishable from the old media. The objective is to refresh the underlying physical media instance without altering the digital information it contains, apart from any media instance identification information.

Representation Information: Information that maps data into more meaningful concepts. An example is the ASCII definition which describes how bits (i.e.,data) are mapped into

numbers. Another example is a description of the numbers (i.e., data) of a table as being the coordinates of a location on the Earth measured in East longitude and latitude.

Representation Net: The set of Representation Information which fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so its digital forms can be understood over the Long Term.

Request Agreement: An agreement between the archive and the Consumer in which the physical details of the delivery such as media type and object format are specified. *[ck after updating section 2 discussion, which now is introduced abruptly - DMS]*

Result Set: The set of descriptive records for those AIPs in an OAIS which match the criteria stated in a Consumer query.

Search Session: A session initiated by the Consumer with the archive during which the Consumer will use the archive finding aids to identify and investigate potential holdings of interest.

Semantic Layer Information: (See Object Layer Information)

Structure Layer Information: Translates the bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

Submission Agreement: An agreement reached between an archive and the Producer that negotiates a data model for the Data Delivery Session. This data model identifies format/contents and the logical constructs used by the Producer and how they are represented on each media delivery or in the telecommunication session. It also transfers legal and physical custody of the data to the archive and the terms defining restrictions and access.

Submission Information Package (SIP): The Information Package identified by the Producer in the Submission Agreement with the OAIS

Subscription Request: A request that is generated by a Consumer for information that is to be delivered periodically on the basis of some event or events.

Transmutation: When moving a digital Information Object to new media, alter its Representation Information, and possibly its associated data object, while attempting to preserve the essential meaning of the Information Object. For example, change an ASCII table to UNICODE and update the representation information accordingly when copying the table to new media.

Unit Descriptor: A type of Package Descriptor that is specialized to provide searchable information about an Archival Informaton Unit.

1.7 REFERENCES

[1] "Preserving Digital Information", Report of the Task Force on Archiving of Digital Information, final report currently available from the URL <http://www.rlg.org/ArchTF/>, May 1, 1996

[2] "Object-Oriented Modeling and Design", by Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W., Prentice Hall, 1991.

[3] "Z39.50 Profile for Access to Digital Collections", currently available at the URL <http://vinca.cnidr.org/protocols/profiles/zdl.html>

2 OAIS CONCEPTS

The purpose of this section is to motivate and describe several key high level OAIS concepts. An extension, and formal modeling of these concepts, is given in Section 4.

The term “archive” has come to be used to refer to a wide variety of storage and preservation functions and systems. Traditionally, an archive is understood as a facility or organization which preserves records, originally generated by or for a government organization, institution, or corporation, for access by public. or private communities. It accomplishes this task by taking ownership of the records, ensuring that they are understandable to the accessing community, and managing them so as to preserve their information content and authenticity. Historically, such records have been in such forms as books, papers, maps, photographs, and film which can be read directly by humans, or read with the aid of simple optical magnification and scanning aids. The major focus for preserving this information has been to ensure that they are on media with long term stability and that access to this media is carefully controlled.

The explosive growth of information in digital forms has posed a severe challenge not only for traditional archives and their information providers, but for many other organizations in the government, commercial and non-profit sectors. These organizations are finding, or will find, that they need to take on the information preservation functions typically associated with traditional archives because digital information is easily lost or corrupted. The pace of technology evolution is causing some hardware and software systems to become obsolete in a matter of a few years, and these changes can put severe pressure on the ability of the related data structures or formats to continue effective representation of the full information desired.

A major purpose of this reference model is to facilitate a much wider understanding of what is required to preserve information for the Long Term. To avoid confusion with simple "bit storage" functions, the reference model defines an Open Archival Information System (OAIS) which performs a Long-term information preservation function. An OAIS archive is one that intends to preserve information for access and use by one or more designated communities, and it meets the minimum requirements given in Section 3. This Archival Information System is composed of personnel and supporting systems, and it is considered to be "Open" because the model and its standards are being developed using a public process that ensures this information is readily available to the public. "Open" does NOT mean that access to information within any archive is uncontrolled. For the remainder of this document, the term archive is understood to refer to an OAIS, or OAIS archive, unless the context makes it clear otherwise (e.g., traditional archives).

The OAIS model recognizes the already highly distributed nature of digital information holdings and the need for local implementations of effective policies and procedures supporting information preservation. This allows, in principle, a wide variety of organizational arrangements, including various roles for traditional archives, in achieving this preservation. It is expected that organizations attempting to preserve information will find that using OAIS terms and concepts will assist them in achieving their information preservation goals.

At the same time, the problems associated with achieving economical and truly effective Long-term digital information preservation should not be underestimated. A good survey of many of the issues is contained in a report by the Task Force on Archiving of Digital Information entitled “Preserving Digital Information” [Reference 1].

2.1 OAIS ENVIRONMENT

The environment surrounding an OAIS is given by the simple model shown in **Figure 2-1**

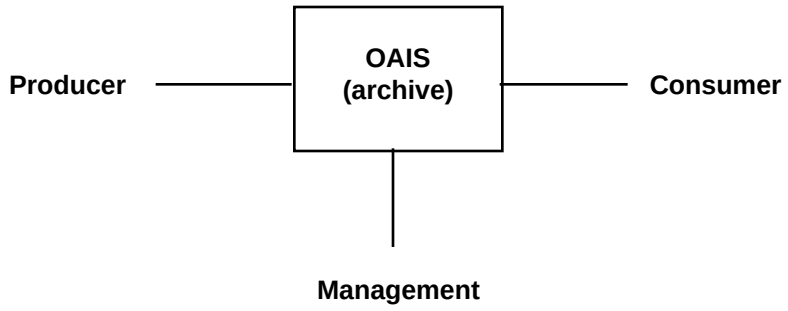


Figure 2-1. Environment Model of an OAIS

Outside the OAIS are **Producers**, **Consumers**, and **Management**. Producer is the role played by those persons, or client systems, who provide the information to be preserved. Management is the role played by those who set overall OAIS policy as one component in a broader policy domain. In other words, Management oversight of the OAIS is only one of Management’s responsibilities. Management is not involved in day-to-day archive operations. The responsibility of managing the OAIS on a day-to-day basis is included within the OAIS in an administrative function and will be described in Section 4. Consumer is the role played by those persons, or client systems, who interact with OAIS services to find and acquire preserved information of interest.

Other OAIS archives are not shown explicitly however such archives may establish particular agreements among themselves consistent with Management and OAIS needs. Other archives may interact with a particular archive for a variety of reasons and with varying degrees of formalism for any pre-arranged agreements. One OAIS may take the role of Producer to another OAIS; an example is when the responsibility for preserving a type of information is to be moved to this other archive. One OAIS may take the role of Consumer to another OAIS; an example is when the first OAIS decides to rely on the other OAIS for a type of information it seldom needs and chooses not to preserve locally. Such reliance should have some formal basis that includes the requirement for communication between the archives of any policy changes which might affect this reliance.

2.2 OAIS INFORMATION

2.2.1 INFORMATION DEFINITION

A clear definition of information is central to the ability of an OAIS to preserve it. While formal modeling of information is given in Section 4, some key concepts are given in this section.

A person, or system, can be said to have a knowledge base which allows them to understand received information. For example, a person whose has a knowledge base that includes an understanding of English will be able to read, and understand, an English text.

Information is defined as any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data. For example, the information in a hardcopy book is typically expressed by the observable characters (the data) on the pages which, when they are combined with a knowledge of the language used (the knowledge base), converts those characters to more meaningful information. If the recipient does not already include English in its knowledge base, then the English text (the data) need to be accompanied with an English dictionary and grammar information (i.e., Representation Information) in a form that is understandable to the recipients knowledge base.

Similarly, the information stored within a CD-ROM file is expressed by the bits (the data) it contains which, when they are combined with the Representation Information for those bits, converts those bits to more meaningful information as long as the Representation Information is understandable to the recipients knowledge base. For example, assume the bits represent an ASCII table of numbers giving the coordinates of a location on the Earth measured in degrees latitude and East longitude. The Representation Information will be the definition of ASCII together with descriptions of the format of the numbers and their locations in the file, their definitions as latitude and longitude, and the definition of their units as degrees.

In general, it can be said that "Data interpreted using its Representation Information yields Information" and this is shown schematically in **Figure 2-2**.

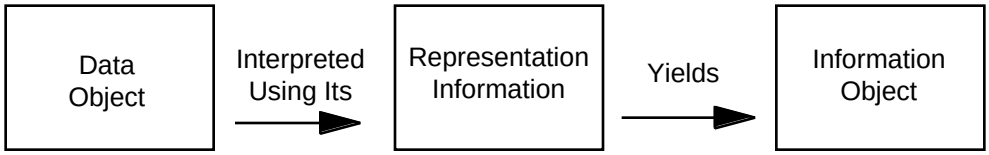


Figure 2-2. Obtaining Information from Data

In order for this **Information Object** to be successfully preserved, it is critical for an archive to clearly identify and understand the **Data Object** and its associated Representation Information. For digital information, this means the archive must clearly identify the bits and the Representation Information that applies to those bits. This required transparency to the bit level is a distinguishing feature of digital information preservation and it runs counter to the information hiding trend which is so successful in supporting modern computing services. This presents a significant challenge to the preservation of digital information. (Note: the recursive nature of Representation Information, which typically is composed of its own data and other Representation Information, is dealt with in Section 4.)

This definition of an Information Object is applicable to all the information types discussed in the following sections. In other words, they all have associated Representation Information although this is usually not shown explicitly.

2.2.2 INFORMATION PACKAGE DEFINITION

Every submission of information to an OAIS by a Producer, and every dissemination of information to a Consumer, occurs as one or more discrete transmissions. Therefore it is convenient to define the concept of an Information Package (IP).

An **Information Package (IP)** is defined to have, in general, three components as shown in **Figure 2-3**. These components are the **Content Information (CI)**, **the Preservation Description Information (PDI)**, and the **Packaging Information (PI)**.

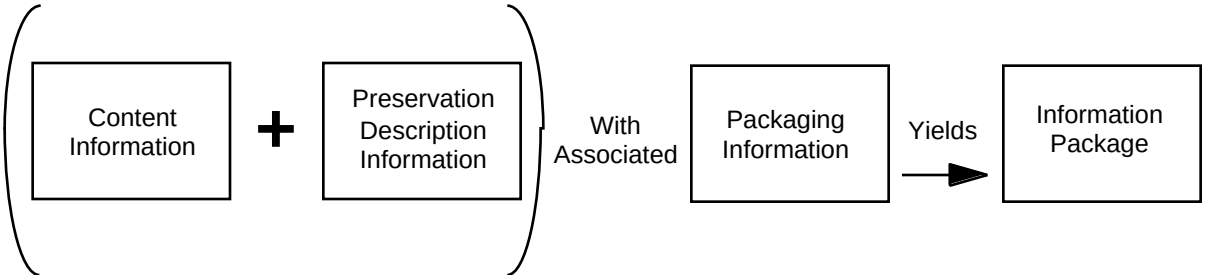


Figure 2-3. Information Package (IP)

The Content Information is that information which is the primary target of preservation. For example, it may be the content of a hardcopy document, or it may be an image provided as the bit content of a CD-ROM file together with the Representation Information for those bits.

Only after the Content Information has been clearly defined can an assessment of the Preservation Description Information be made. The Preservation Description Information applies to the Content Information and is needed to make the Content Information preservable over the Long Term. The Preservation Description Information is broken down into four types of preserving information called Provenance, Context, Reference, and Fixity. Briefly, they are the following:

- Provenance describes the source of the CI, who has had custody of the CI since its origination, and what its history (including processing history) has been.
- Context describes how the CI relates to other information outside the Information Package. For example, it would describe why the CI was produced and it may include a description of how it relates to another CI object that is available. It does not, however, describe how the CI is organized on some storage media. This task is handled by the Packaging Information which tends to be much more volatile than the

CI and PDI.

- Reference provides one or more identifiers, or systems of identifiers, by which the CI may be uniquely identified. An ISBN number for a book is one example.
- Fixity provides a wrapper, or protective shield, that protects the information from unintended alteration. For example, it may involve a check sum over the CI of a digital Information Package.

The Packaging Information is that information which, either actually or logically, binds and relates the components of the package into an entity. For example, if the CI and PDI are identified as being the content of specific files on a CD-ROM, then the Packaging Information may be the ISO-9660 volume/file structure on the CD-ROM. The Packaging Information, in this case, may also include the physical CD-ROM disk. These choices are the subject of local archive definitions or conventions. The Packaging Information may be altered, particularly during migrations to new media types, and its data are not a part of the information that must be preserved.

In concluding this section, the importance of first identifying the Content Information, and then assessing the Preservation Description Information that applies to it, can not be over-estimated when preparing and preserving Information Packages.

2.2.3 INFORMATION PACKAGE VARIANTS

It is necessary to distinguish between an Information Package that is preserved by an OAIS and the Information Packages that are submitted to, and disseminated from, an OAIS. These variant packages are needed to reflect the reality that some submissions to an OAIS will have insufficient Representation Information or PDI to meet final OAIS preservation requirements. In addition, they may be organized very differently from the way the OAIS organizes the information it is preserving. Finally, the OAIS may provide information to Consumers that does not include all the Representation Information or all the PDI with the associated Content Information being disseminated. These variants are referred to as the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP). Although they are all information packages, they differ in what is mandatory content.

The **Submission Information Package (SIP)** is that package that is sent to an OAIS by a Producer. Its form and detailed content is typically negotiated between the Producer and the OAIS. Most SIPs will have some CI and some PDI, but it may require several SIPs to provide a complete set of CI and associated PDI. The PI will always be present in some form.

Within the OAIS one or more SIPs is transformed into one or more **Archival Information Packages (AIP)** for preservation. The AIP has a complete set of PDI for the associated CI. The AIP may also contain a collection of other AIPs and this is discussed and modeled in Section 4. The Packaging Information of the AIP will conform to OAIS internal standards, and it may vary as it is managed by the OAIS.

In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a **Dissemination Information Package (DIP)**. The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will always be present in some form so that the Consumer can clearly distinguish the information requested. The Packaging Information may take several forms depending on the dissemination media and Consumer requirements.

2.3 OAIS HIGH LEVEL EXTERNAL INTERACTIONS

The following sections present a high level view of the interaction between the entities identified in the OAIS environment. **Figure 2-5** is a data flow diagram that represents the operational OAIS archive external data flows. This diagram concentrates on the flow of information among Producers, Consumers and the OAIS and does not include flows that involve Management. These flows are dealt with further in Section 4.

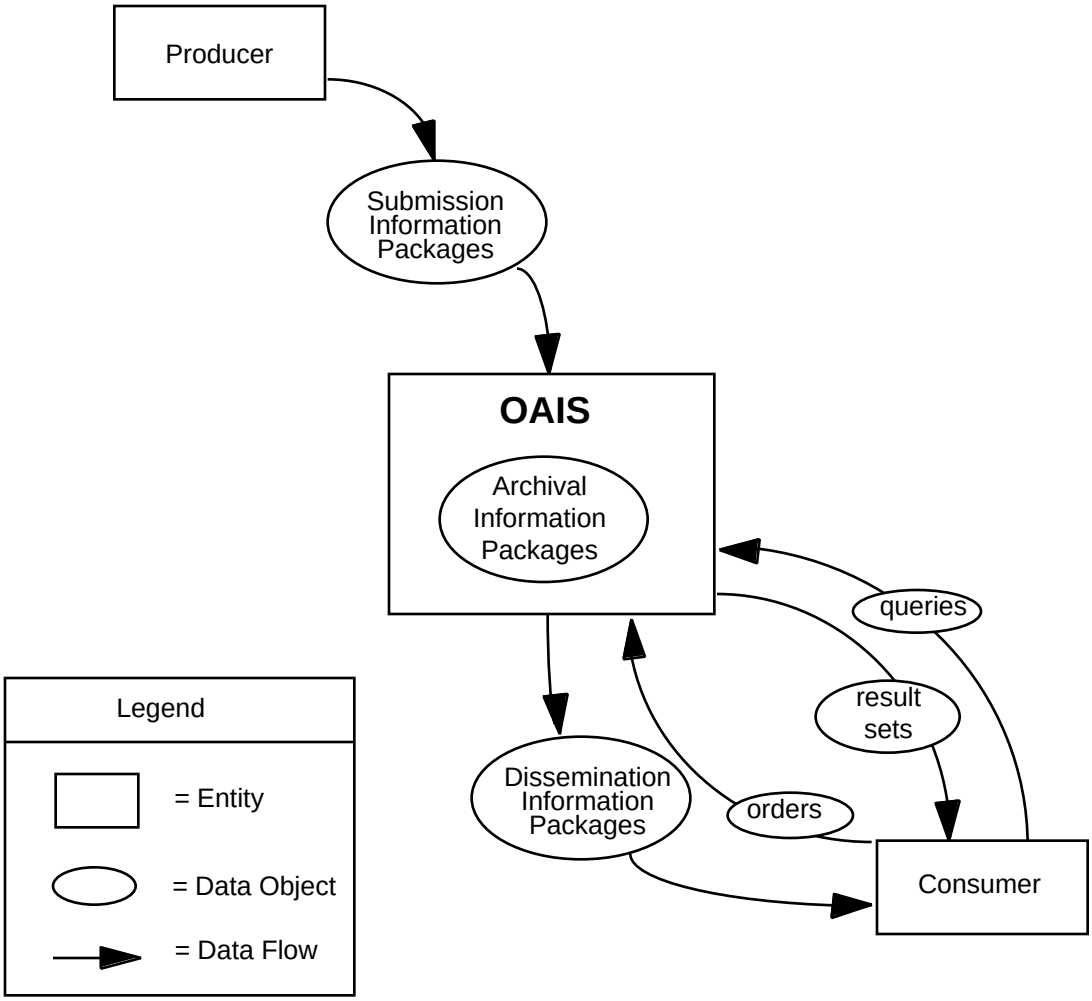


Figure 2-5. OAIS Archive External Data Flows

2.3.1 PRODUCER INTERACTION

The Producer establishes a **Submission Agreement** with the OAIS, which identifies the SIPs to be submitted and may span any length of time for this submission. Some Submission Agreements will reflect a mandatory requirement to provide information to the OAIS, while others will reflect a voluntary offering of information. Within the Submission Agreement, one or more **Data Delivery Sessions**, usually with significant time gaps between the sessions, are recognized. A Data Delivery Session will contain one or more SIPs and may be a delivered set of media or a single telecommunications session. The Data Delivery Session content is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical components of the SIP (e.g., the CI, PDI, and PI) that are to be provided and how they are represented on each media delivery or in the telecommunication session. All data deliveries within a Submission Agreement are recognized as belonging to that Submission Agreement and will generally have a consistent data model which is specified in the Submission Agreement. For example, a Data Delivery Session might consist of a set of CI objects corresponding to a set of observations which are carried by a set of files on a CD-ROM. The names of the files or the directory structure of the CD-ROM might be used to store information about the observation times or the datatypes contained in each file. The Submission Agreement would indicate how the file format and the convention for the filenames is to be provided, and it would give the frequency of Data Delivery Sessions (e.g. one per month for two years). It would also give other needed information such as access restrictions to the data and it would state how the PDI is to be provided if it is not a regular part of each SIP.

Each SIP in a Data Delivery Session is expected to meet minimum OAIS requirements for completeness. However multiple SIPs may need to be received before an acceptable AIP is formed and fully ingested within the OAIS. At this point the information contained therein can become available, in principle, to OAIS Consumers. A Submission Agreement also includes, or references, the procedures and protocols by which an OAIS will either verify the arrival and completeness of a Data Delivery Session with the Producer or question the Producer on the contents of the Data Delivery Session.

2.3.2 CONSUMER INTERACTION

There are two basic request types initiated by Consumers, the **Subscription Request** and the **Adhoc Request**.

– Subscription Request: The Consumer establishes a **Request Agreement** with the OAIS for information expected to be received periodically on the basis of some triggering event. It may span any length of time and under it one or more **Data Dissemination Sessions**, usually with significant time gaps between the sessions, may take place. A Data Dissemination session may involve the transfer of a set of media or a single telecommunications session. The Request Agreement identifies the components of one or more AIPs to be provided, identifies how they are mapped into a Dissemination Information Package (DIP) and how that DIP will be packaged in each media delivery or telecommunications session. The Request Agreement will also specify other needed

information such as the trigger (e.g. event or time period) for new Data Dissemination sessions, the criteria for selecting the OAIS holdings to be included in each new Data Dissemination session, delivery information (e.g., name or mailing address), and any pricing agreements.

– Adhoc Request: The Consumer establishes a Request Agreement with the OAIS for information available from the archive. If the Consumer does not know a priori what specific holdings of the OAIS he is interested in acquiring, the Consumer will establish a **Search Session** with the OAIS. During this Search Session the Consumer will use the OAIS finding aids, which operate on catalogued information or, in some cases on the AIPs themselves, to identify and investigate potential holdings of interest. This may be accomplished by the submission of queries and the return of result sets to the Consumer. This searching process tends to be iterative with a user first identifying broad criteria and then refining this criteria based on previous search results. Once the Consumer identifies the OAIS AIPs, or AIP components, he wishes to acquire, he must provide a Request Agreement to the OAIS to document the details of what components he will acquire and how he will acquire them. At this point the Adhoc Request and the Subscription Request are similar although the Request Agreement may be substantially simpler with the Adhoc Request. Note that the concept of a Request Agreement does not specify any particular implementation. It may, in some cases, be no more than the completion of a World Wide Web form.

2.3.3 MANAGEMENT INTERACTION

– Management provides the OAIS with its charter and scope. The charter may be developed by the archive but it is important that Management formally endorse archive activities. The scope determines the breadth of both the Producer and Consumer groups served by the archive.

– Management is often the primary source of funding for an OAIS and may provide guidelines for resource utilization (personnel, equipment, facilities).

– Management will generally conduct some regular review process to evaluate OAIS performance and progress toward long-term goals.

– Management determines or at least endorses pricing policies for OAIS services.

– Management participates in conflict resolution involving Producers, Consumers and OAIS internal administration.

– Management should also provide support for the OAIS by establishing procedures that assure OAIS utilization within its sphere of influence. For example, management policies should require that all funded activities within its sphere of influence submit data products to the archive and also adhere to archive standards and procedures.

3 OAIS RESPONSIBILITIES

This section establishes mandatory responsibilities that an organization must discharge in order to operate an OAIS archive. The OAIS must:

- Negotiate and accept information from information producers.

- Determine (dependently or independently) which communities need to be able to understand the information provided.

- Ensure the information to be preserved is independently understandable to the designated communities. In other words, the communities should be able to understand the information without needing the assistance of the experts who produced the information.

- Assume sufficient control of the information provided to the level needed to ensure Long-term preservation.

- Follow documented policies and procedures which ensures the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.

- Make the preserved information available to the designated communities in forms understandable to those communities.

In the following sections, each of these responsibilities is explored in greater depth for clarification.

3.1 NEGOTIATES AND ACCEPTS INFORMATION

An organization operating an OAIS will have established some criteria that aids in determining the types of information that it is willing to, or it is required to, accept. These criteria may include, among others, subject matter, information source, degree of uniqueness or originality, and the nature of the techniques used to represent the information (e.g., physical media, digital media, format). The information may, in general, be submitted using a wide variety of common and not-so-common forms, such as books, documents, maps, data sets, and moon rocks using a variety of communication paths including networks, mail, and special delivery.

To extract some commonalty across these forms, the reference model has defined the concept of an "Information Package." The IP provides a conceptual basis around which to formulate more specific archival policies on what is acceptable in a Submisison Information Package (SIP). It requires that a clear distinction be made about what constitutes the primary information that is to be preserved (i.e., Content Information) so that the Preserving Description Information can be clearly defined and assessed in relation to the Content Information. The Content Information, for a digital submission, is identified by the bits that

are assigned to it. These bits include any bits used to express Representation Information needed as part of the Content Information. In general, many SIPs may be needed to complete one Archival Information Package (AIP) or one SIP may result in many AIPs. Therefore some SIPs may only contain Content Information and others may only contain Preserving Description Information, in addition to their Packaging Information. The relationships among SIPs, which are all part of the same Submission Agreement, needs to be made clear in the submission process.

Ideally, the OAIS and the data Producer who is submitting information to the OAIS, will agree in advance on what the SIPs should be and how they will be delivered as documented in the Submission Agreement. This should include all the information the OAIS needs to form one or more AIPs. This will facilitate the ingest of this information into the OAIS. However, the OAIS may need to redefine the form of the AIPs, corresponding to a set of provided SIP information, in order to provide effective consumer services. This may also include the need to increase or decrease the granularity of information which Consumers may request from the OAIS. It may also include the creation of 'derived AIPs' which are generated from stored algorithms run against stored AIPs at the time of Consumer requests. The OAIS may provide Consumers with one, or many, model views of its holdings.

For example, a ten-page manuscript may be submitted as the Content Information of a SIP. If its Preserving Description Information is adequate, then there is a direct mapping to an AIP and it may be stored and catalogued as such for Consumer request. Subsequently, each page may be made independently locatable to Consumers using an OAIS-generated index. If each page is also documented with appropriate Preserving Description Information, then each page effectively becomes the Content Information of new AIP as well. The manuscript as a whole remains as Content Information of an aggregating AIP. This aggregation of types of AIPs is modeled in Section 4.

3.2 DETERMINES DESIGNATED CONSUMER COMMUNITIES

The submission, or planned submission, of a SIP requires a determination as to who the expected consumers of this information will be. This is necessary in order to determine if the information, as represented, will be understandable to that community.

For example, an archive may decide that a SIP's Content Information should be understandable to the general public. It will need to be sure that the resulting AIPs Content Information and Preserving Description Information, expected to be widely understandable to the general public, is free of jargon. For digital information, this means that the associated Representation Information must also be free of jargon and widely understandable. These concepts, and the role of software, are discussed more fully in Sections 3.3, 3.4, and 4.

For some scientific information, the designated community of consumers might be described as those with a first year graduate level education in a related scientific discipline. This is a more difficult case as it is less clear what degree of specialized scientific terminology might actually be acceptable. The producers of such specialized information are often familiar with a narrowly recognized set of jargon, so it is especially critical to clearly define the designated

community for their information and to make the effort to ensure this community can understand the information.

The possible evolution of the designated community also needs consideration. Information originally intended for a narrowly defined community may need to be made more widely understandable at some future date. This is likely to mean adding additional explanations in support of the Representation Information and the Preserving Description Information and it can become increasingly difficult to obtain this information over time. Selecting a broader definition of the designated community when the information is first proposed for Long-term Preservation can reduce this concern and also improve the likelihood that the information will be understandable to all in the original community.

3.3 ENSURES INFORMATION IS INDEPENDENTLY USABLE

The degree to which an AIP conveys information to a designated community is, in general, quite subjective. Nevertheless, it is essential that an archive make this determination in order to maximize information preservation.

For example, a manuscript's Content Information may be written in English and therefore its content may be generally understandable to a wide audience. However, unless the purpose for which it was created is clearly documented, much of its meaning may be lost. This 'purpose' information is part of its Context and must be provided in the Preserving Description Information.

As another example, consider AIPs from a digital set of observations of rainfall, temperature, pressure, wind velocities, and other parameters measured all over the world for a year. This type of information is very extensive, is not usually in a form intended for direct human browsing or reading, but it is in a form appropriate to searching and manipulation by application software. Such content may only be understandable to the original producers unless there is adequate documentation of the meaning of the various fields and their inter-relationships, and how the values relate back to the original instrumentation that made the observations. In such specialized fields extra effort is needed to ensure that the Content Information and the Preservation Description Information are understandable to a designated community. If the archive does not have this level of expertise in-house, it may need to have outside community representatives review the information for long-term understandability. Otherwise the information may be understandable to only a few specialists and be lost when they are no longer available.

Digital Content Information needs software for efficient access. However, maintaining Content Information-specific software over the long term has not yet been proven cost effective.

In general, each Content Information object should have its Representation Information fully documented down to the bit level, even if the lower levels of Representation Information are, at ingest time, supported by widely available software. Over time, all representations will be replaced and documenting them ensures that the OAIS can track them as to viability for

usage by the OAIS Consumers. This also facilitates Content Information migration to new representations when needed.

3.4 ASSUMES SUFFICIENT CONTROL FOR PRESERVATION

In general, the OAIS will accept the SIPs as either a custodian or as the new legal owner. When acting as a custodian, the OAIS may need to involve the actual owner(s) in some migration and access decisions depending on the authority it has been granted to act independently. When it is the legal owner, it already has the independence to do what is needed to preserve the information and make it accessible.

Upon acceptance of SIPs and formation of an AIP, the OAIS must assume sufficient control over the Content Information and Preservation Description Information so that it is able to preserve it for the Long Term. There is no issue with the AIP's Packaging Information because, by definition, this is under internal OAIS control. The problems of assuming sufficient control of the Content Information and Preservation Descriptiopn Information, which are largely digital, are addressed in four related categories as follows:

- Obtaining complete Representation Information
- Authority to modify Representation Information
- Agreements with external organizations
- Copyright implications

Obtaining Complete Representation Information: For an AIP in the possession of the OAIS, the presence of Fixity information in the Preservation Description Information is to assure that the Content Information bits and the Content Information's Representation Information are not altered. However determining what is sufficient Representation Information is not always straight forward. For example, a scientific data set may be a file containing several types of data, including text and images and tables. The formats and meanings of each of these components may be described in separate documents, and the overall data file structure may be described in yet another document. Each of these documents may be in an electronic form which means they each have additional Representation Information that is needed to understand them. These recursions only end when the Representation Information objects (documents) are expressed in a hardcopy (non-digital) form that is readable and understandable.

Ideally, an OAIS would include all the digital Representation Information objects associated with a Content Information in its AIP, and it would point to the hardcopy documents needed to complete the description chains. Further, these hardcopy documents would also be in the OAIS. While this happens some of the time, often the representation chains are broken when widely available software is used in place of obtaining a description. If, in the example above, the images are in the Graphics Interchange Format (GIF), the GIF Representation Information may not be included in the AIP and may only be referenced because it is assumed that GIF display software will be available to present the images. **These assumptions need to be explicitly recorded in the OAIS so they can be periodically checked because eventually they will not be true.** At some point a decision will have to be

made to either include the missing GIF Representation Information, or the images will have to be migrated to a new image format and a decision made on whether this new format should be described and/or just referenced because appropriate software is widely available.

Authority to modify Representation Information: Although the Fixity information within the Preserving Description Information of an AIP is ensuring that the Content Information related bits are not being altered, there will come a time when Content Information bits are not in a form that is convenient for the designated Consumer community. The Content Information bits may be fully documented in available hardcopy forms, so technically the information has not been lost, but practically the information has become inaccessible. The OAIS needs the authority to migrate the Content Information to new representation forms. It may have to bring in subject matter experts, from outside the OAIS, to help ensure that information is not lost. Ideally, when this situation arises, both the original AIPs (fully described) and new AIPs will be retained. Migration issues are addressed more fully in Section 5.

Agreements with external organizations: An OAIS may establish a variety of agreements with other organization to assist in its preservation objectives. It may establish an agreement with another OAIS so that it does not have to preserve all the Representation Information objects related to its Content Information objects. For example, one OAIS may hold a description of GIF that can be referenced by many other OAIS archives. This begins to form a Federation of archives to achieve common purposes more effectively than could be done independently. This will work only when there is a sufficient level of trust among these archives. One way to enhance this trust is the use of archive certification programs.

Agreements with non-OAIS organizations could be useful in helping to acquire and ingest high quality SIPS that are well on their way to becoming AIPs. However an OAIS can not be responsible for information held by such organizations and such information can not be said to be under ‘effective preservation for the Long Term’.

When the OAIS acts as a custodian, it will have to abide by the agreements it has reached with the owner of the information. If it feels the information is in danger of becomming lost, and it does not have sufficient control to make needed transformations, it will need to renegotiate its agreement.

Copyright implications: These issues obtain when the OAIS acts as a custodian. An OAIS that acquires copyrighted material has special concerns to address. Copyright issues surrounding digital materials are in a state of flux. Nevertheless, it seems clear that an OAIS will have to be an active player is supporting copyright rules and thus will need to control access to some of these materials. Particularly tricky will be the migration issues when the Content Information needs to be transformed. It will be important to establish just what has been copyrighted and the extent to which the Content Information bits can be altered.

3.5 FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES

It is essential for an OAIS to have documented policies and procedures for preserving its AIPs, and to follow those procedures. This is particularly true for digital AIPs or digital Content Information objects because of their frequent need for attention as discussed earlier. They are also highly vulnerable to inadvertent and intentional destruction or corruption even in the most trusted systems. This fragile nature puts a premium on being able to recover from all digital AIP handling operations when errors are discovered.

The appropriate policies and procedures will depend, at minimum, on the nature of the AIPs and any 'backup' relationships the archive may have with other archives.

In general, digital SIPs received may undergo some transformations before being fully incorporated into the OAIS as AIPs, may undergo migrations that include transformations while in the OAIS, and may undergo transformations upon dissemination as DIPs to Consumers. For example, one extreme on ingest is to extract all relevant 'values' from the Content Information objects (thus forming new Content Information objects) and to store them according to a complex schema, or data model, covering the archive's subjects of interest. Such a transformation should be fully documented and traceable to the original SIPs. During internal migrations, some or all of the representations used to carry the information of an AIP, including its Packaging Information, may be replaced. These new representation need to be carefully documented and the transformation process needs to be fully described. Upon dissemination of an AIP, or part of an AIP, to a Consumer, a new representation may be used to provide more effective usage. Again, this transformation needs to be fully described and the descriptions of all past transformations need to be available to the Consumer. This attention to detail, while also ensuring against processing errors, requires that strong policies and procedures be in place and that they be executed.

A long-term technology usage plan, updated as technology evolves, is also essential to avoid being caught with very costly system maintenance, emergency system replacements, and costly data representation transformations.

3.6 MAKES THE INFORMATION AVAILABLE

By definition, an OAIS makes its AIPs visible and available to Consumers. Multiple views, supported by various search aids that cut across collections of AIPs, may be provided. Some AIPs, visible to Consumers, may not be stored as AIPs in any recognizable sense. Rather, they are 'derived AIPs' that may be generated upon request using an associated algorithm operating on one or more existing AIPs. Upon dissemination, these derived AIPs will need to include documentation on how they were derived from other AIPs. The expectations of OAIS Consumers regarding access services will vary widely among archives and over time as technology evolves. Pressures for ever more effective access must be balanced with the requirements for preservation under the available resource constraints.

For security, direct access by Consumers to stored AIPs should only be allowed on copies of the original AIPs. Some AIPs may have restricted access that will require approval of the Consumer before the AIP, or its parts, are disseminated. The OAIS needs to have published policies on access and restrictions so that the rights of all parties are protected.

In general, AIPs will be distributed by all varieties of communication paths, including networks and physical media.

4 DETAILED MODELS

The purpose of this section is to provide a more detailed model view of the functional entities of the OAIS and the information handled by the OAIS. This aids OAIS designers of future systems and provides a more precise set of terms and concepts for discussion of current systems.

4.1 FUNCTIONAL MODEL

The OAIS of Figure 2-1 is broken into six functional entities and related interfaces as shown in **Figure 4-1**. The lines connecting entities identify communication paths over which information flows in both directions. The lines shown to Administration are dashed only to reduce diagram clutter.

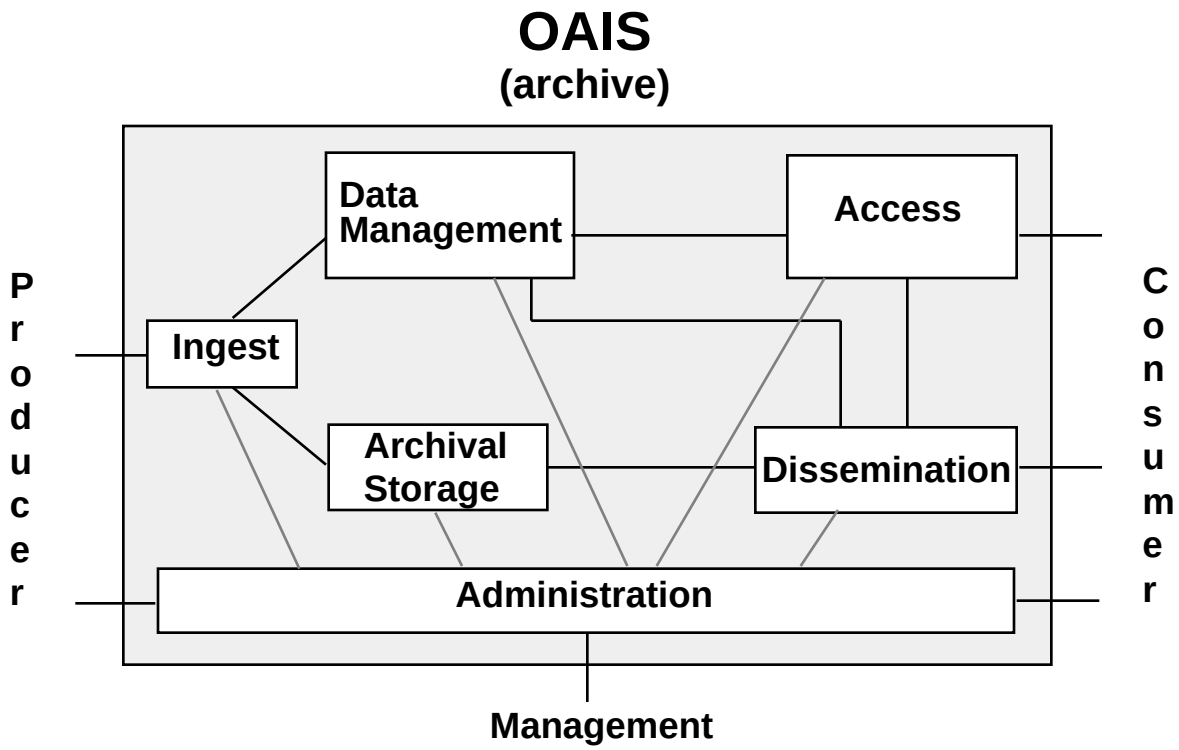


Figure 4-1. OAIS Functional Entities

The role provided by each of the entities in Figure 4-1 is briefly described as follows:

- **Ingest:** This entity provides the services and functions to accept and validate a Submission Information Package (SIP) from Producers and prepare the contents for storage and management within the archive. In summary there is a scheduling function to negotiate a submission agreement and the delivery of one or more SIPs; each SIP is physically received by the staging function; the Content Information (CI) may go through

a conversion process to comply with the internal archive data model and descriptive information may be added to enhance the utility of the package and result in an Archival Information Package (AIP); the AIP is reviewed by the archive staff and others; **Descriptive Information** is derived from the AIP for cataloging in the data management system; and the AIP is made available for transfer to Archival Storage.

- **Archival Storage:** This entity provides the services and functions for the storage and retrieval of AIPs. Archival Storage functions include receiving AIPs from staging storage to permanent Archival Storage; managing the Archival Storage hierarchy; physically migrating data to new media over time; performing routine and special error checking; providing backup procedures; and providing access to AIPs for dissemination.

- **Data Management:** This entity provides the services and functions for populating, maintaining, and querying a wide variety of descriptive information such as catalogs, processing histories and processing algorithms. It also provides Ingest, Access, Dissemination and Administration with system information (e.g. consumer access information, security information, and operational schedules). Data Management functions include providing services for requesting and generating reports; providing the capability for both transactional updates (loading new descriptive information or archive operational statistics) to the data base and periodic review updates; and general data base administration functions (maintaining schema and view definitions and referential integrity). This entity, together with Archival Storage, conceptually contains all the persistent information needed for OAIS operations.

- **Administration:** This entity manages all of the system activities. The Administration functions include planning and scheduling archive facility resources, maintaining configuration management of system hardware and software, performing accounting functions to bill consumers for services; providing service functions to consumers; performing data engineering work to develop and maintain archive standards and policies; and providing a point of interaction with OAIS Management.

- **Access:** This entity supports consumers in determining the existence, description, location and availability of information stored in the OAIS. The Access entity includes functions for providing a mechanism (access session) for the Consumer to communicate with the OAIS; applying access controls to this communication interface; allowing the Consumer to peruse finding aids; and for selecting data objects for on-line or off-line dissemination.

- **Dissemination:** This entity includes the functions which provide archive products to Consumers. The Dissemination entity includes functions for receiving dissemination requests from the Access entity; interacting with Archival Storage to retrieve requested AIPs; generating any necessary descriptive information required to accompany the Dissemination Information Packages (DIP); performing processing that is required to convert the AIP into a DIP; making the resulting DIPs available on-line or off-line; and monitoring the status of the dissemination request until successful completion.

In addition to the entities described above, there are various common services assumed to be available, and all of these are discussed in more detail in Sections 4.1.1 through 4.1.7. Specific flows of information among the entities are shown here in bold type, followed by an identifier in braces. For example, Section 4.1.3 identifies the flow "**storage confirmation {4.1.3g}.**" This and all other flows are then illustrated in the set of data flow diagrams collected in Section 4.1.8.

4.1.1 COMMON SERVICES

Modern, distributed computing applications assume a number of supporting services such as inter-process communication, name services, temporary storage allocation, exception handling, security, and directory services. Common Services provides a single conceptual source for these services.

4.1.2 INGEST

The functions of the Ingest entity are detailed below:

* The **Scheduling** function develops a **submission agreement {4.1.2c}** and negotiates a **data submission schedule {4.1.2k}** with the producer. It also maintains a calendar of expected Data Delivery Sessions that will be needed to transfer one or more complete Archive Information Packages to the OAIS and the resource requirements to support their ingestion.

* The **Staging** function provides the appropriate storage capacity or devices to receive a **SIP {4.1.2a}** from the producer. The SIPs may be delivered via electronic transfer (e.g. ftp); loaded from media submitted to the archive; or simply mounted (e.g. CD-ROM) on the archive file system for access. The Staging function may represent a legal transfer of custody for the CI in the SIP, and may require that special access controls be placed on the contents. The Staging function provides a **confirmation of receipt {4.1.2f}** of a SIP to the Producer.

* The **Conversion** function transforms one or more SIPs into one or more AIPs that conforms to the internal data model of the archive. This may involve file format conversions, data representation conversions or reorganization of the content information in the SIPs.

* The **Review** function provides a validation of the SIP (for a partial ingestion) or the AIP after it has been converted to conform to the internal data model of the archive. The Review function may be carried out on a single SIP or may be withheld until a number of Data Delivery Sessions have been concluded and all the components of the AIP have been accumulated. The review is carried out by the archive data engineers and may also involve an outside committee (e.g., peer review). The review process must verify that the SIP has been physically transported correctly to the archive staging area and successfully converted to internal archive format; that the quality of the data meets the requirements of the archive and the review committee; that there is adequate Representation Information and Preservation Description Information to ensure the Content Information is understandable

and independently usable to the Designated Community; and that the Descriptive Information is sufficient to make the preserved information adequately findable by Consumers from the Designated Community. The formality of the review will vary depending on internal archive policies. The review process may determine that some portions of the SIP are not appropriate for inclusion in the archive and must be resubmitted or excluded. After the review process is completed any **liens {4.1.2b}** are reported to the producer who will then **resubmit {4.1.2i}** or **appeal {4.1.2j}** the decision. After the review is completed a **final ingest report {4.1.2g}** on the Data Delivery Session is prepared for distribution to Administration and to the Producer.

* The **Transfer Initiation** function starts the process that moves the AIP from the staging area to the storage area. This may be either an electronic, physical, or a virtual (i.e. data stays in place) transfer. Transfer of the **AIP {4.1.2p}** includes a **storage request {4.1.2h}**. After completing the transfer, Archival Storage returns a **storage confirmation {4.1.2m}** indicating the storage location of the AIP. This confirmation is included in the database update prepared by the Cataloging function.

* The **Cataloging** function extracts descriptive components of the SIPs or completed AIPs to populate the data management system. This **descriptive information {4.1.2n}** consists of selected parameters that are needed to support the functions of the Access and Dissemination entities. The transfer of descriptive information to the Data Management entity includes a **database update request {4.1.2d}.** In return, Data Management provides a **database update response {4.1.2e}** indicating the status of the update.

4.1.3 ARCHIVAL STORAGE

The functions of the Archival Storage entity are detailed below:

* The **Transfer Receiving** function receives a **transfer request {4.1.3d}** of an archival information package from staging Archival Storage and moves the data to permanent Archival Storage within the archive. The transfer request may need to indicate the anticipated frequency of utilization of the data objects comprising the Archival Information Package (AIP) to allow the appropriate storage devices or media to be selected for storing the AIP. The Transfer Receiving function will select the media type, prepare the devices or volumes and perform the physical transfer to the Archival Storage volumes. On completion of the transfer, this function sends an **Archival Storage confirmation {4.1.3g}** message to Ingest.

* The **Hierarchy Management** function positions the AIPs (AIUs and AICs) on the appropriate media based on directions from ingest (transfer request), administrative **policies {4.1.3h}** or usage statistics. This function provides regular reports to Administration summarizing the **inventory of media on-hand {4.1.3a}**, **available storage capacity {4.1.3b}** in the various tiers of the storage hierarchy, and other **operational statistics {4.1.3c}**.

* The **Physical Migration** function provides the capability to reproduce the AIPs over time. Refer to Section 5 for a detailed description of migration issues. The fundamental rule of

data migration under the Physical Migration function is that the Content Information and Preservation Description Information must not be altered. However the data constituiting the Packaging Information may be changed as long as it continues to perform the same function. The migration strategy must take into consideration the expected and actual rates of errors encountered in various media types, their performance, and the costs of ownership when deciding what media to migrate to. If media dependent attributes (e.g. tape block sizes, CD-ROM volume information) have been included as part of the Content Information, a way must be found to preserve this information when migrating to higher capacity media with different storage architectures.

* The **Error Checking** function provides statistically acceptable assurance that no components of the archive information package are corrupted during transfer receiving, migration, backup or duplication procedures. This function requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard logs that are checked by the Archival Storage staff. The PDI Fixity Information provides some assurance that the Content Information has not been altered as the AIP is moved and acessed. Similar information is needed to protect the PDI itself. A standard mechanism for tracking and verifying the validity of all data objects within the archive could also be used. For example, cyclical redundancy checks (CRCs) could be maintained for every individual data file. The storage facility procedures should provide for random verification of the integrity of data objects using CRCs or some other error checking mechanism.

* The **Disaster Recovery** function provides a mechanism for producing duplicate copies of AIPs (AIUs and AICs) in the archive collection. The backup media should be capable of being removed from the archive for storage at a separate facility. **Disaster Recovery policies {4.1.3i}** are specified by Administration.

* The **Provide Data** function provides copies of stored AIPs to dissemination. The function receives a **data request {4.1.3j}** from dissemination which identifies the requested **AIP(s) {4.1.3m}** and either the output media type or a staging area for electronic transfers. The Provide Data function sends a **notice of data transfer {4.1.3k}** to Dissemination.

4.1.4 DATA MANAGEMENT

The functions of the Data Management entity are detailed below:

* The **Report Generation** function receives a **report request {4.1.4b}** from Ingest, Access, Administration or Dissemination and prepares the necessary queries to generate the report.

* The **Report Production** function sends the **report {4.1.4a}** to the requester. It also provides the capability to store report requests and to generate periodic reports or reports triggered by logical criteria on a periodic basis.

* The **Update** function adds, modifies or deletes information in the Data Management persistent storage. There are three major sources of updates; ingest transactions, system

updates, and review updates. Ingest transactions identify new AIPs stored in the archive {flow from Ingest}. System updates include all system related information (Consumer information, request tracking). Review updates are generated by periodic reviewing and updating of information values (e.g. contact names, and addresses). The update function provides regular reports to administration summarizing the **status of updates to the data base {4.1.4e}**.

* The **Data Base Administration** function is responsible for maintaining the integrity of the Data Management persistent storage; for creating any schema or table definitions required to support data management functions; and for providing the capability to create, maintain and access customized user views of the contents of this storage.

4.1.5 ADMINISTRATION

The functions of the Administration entity are detailed below:

* The **Planning and Scheduling** function schedules system usage. It keeps records on times of heavy resource utilization, system down times for maintenance, and system upgrades. The planning and scheduling function also **solicits desirable archivable information {4.1.5b}** for inclusion into the OAIS and handles administrative aspects of acquiring new SIPs.

* The **Configuration Management** function maintains configuration control over the archive system, systematically controlling changes to the configuration. This function maintains integrity and tractability of the configuration during all phases of the system life cycle. It also audits system operations, system performance, and system usage.

* The **Physical Access Control** function provides mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the archive as determined by archive policies.

* The **Customer Service** function provides any necessary assistance to archive system consumers. This service will include **answering questions, resolving consumer problems {4.1.5d}**, providing information about and **documentation {4.1.5e}** on the system, providing status of orders, and providing information about the **status of data ingest {4.1.5g}** activities. The Customer Service function will also create, maintain and delete consumer accounts. It will **bill {4.1.5a}** and collect **payment {4.1.5c}** from consumers for the utilization of archive system resources.

* The **Data Engineering** function supports all Ingest functions and is responsible for developing and maintaining the archive system data standards. The data submission formats and procedures must be clearly documented in the archive's data submission manual and the deliverables must be identified by the producer in the submission agreement. It will also develop **policies {4.1.3h}** for Archival Storage hierarchy management.

* The **Management Interaction** function receives and carries out Management policies. These policies include the **OAIS charter {4.1.5g}**, **scope {4.1.5h}**, **resource utilization guidelines {4.1.5i}**, and **pricing policies (3.1.5j}**. It also provides OAIS **performance**

information {4.1.5k} to Management.

4.1.6 ACCESS

The functions of the Access entity are detailed below:

* The **Provide Access Session** function provides a user interface to the information holdings of the archive. This interface will normally be via computer network or dial up link to an on line service, but might also be implemented in the form of a printed catalog, or fax-back type service. This function also provides a hierarchy of security controls depending on the needs of the archive system. These include establishing firewalls to prevent communication outside an area, electronic signatures and authorization procedures, restricting access to certain network domains, and assignment of user names and passwords. Any combination of these procedures may be needed in certain archive scenarios.

* The **Prepare Finding Aids** function provides tools and products which provide an overview of AIUs and AICs available in the archive system. Finding aids include summary versions of products which can be quickly viewed such as thumbnails images, or abstracts of documents. This function also generates **requests** for specialized queries or processing functions to be carried out by Dissemination to produce new representations of the data objects to extend the retrieval capabilities of the Data Management function (e.g., data mining).

* The **Accept Dissemination Request** function accept a **dissemination request {4.1.6b}** from a user, insure its validity, verify that all required information has been provided, and prepare the request for execution by the Dissemination entity. The dissemination request may be a **subscription request {4.1.6c}** or **adhoc request {4.1.6d}**. This function will provide the Consumer the opportunity to review and correct the information in the request. It will also provides the capability to request special processing of data prior to dissemination via the process data function.

4.1.7 DISSEMINATION

The functions of the Dissemination entity are detailed below:

* The **Receive Dissemination Request** function accepts a **dissemination request {4.1.7a}**. A unique order number/identification is assigned to each accepted dissemination request. For each request accepted, this function adds an entry to the pending orders reflecting that this accepted order has not yet been filled. All order information is verified, and if any errors or unusual conditions are found the Consumer will be notified via an **error message {4.1.7k}**.

* The **Generate DIP** function accepts a dissemination request, validates the request using the **package descriptors {4.1.7j}**, **retrieves the data {4.1.7m}** from Archival Storage and moves a copy of the data to a staging area for further processing. This function also transmits a **report request {4.1.4b}** to Data Management. This function accepts and validates the commands, stores the validated selection parameters (field values) from the

command, and then performs the necessary retrieval operations to find and access the requested data. It also accesses Data Management to obtain consumer information such as consumer's name, address, account number, preferred distribution method or media, and other consumer-oriented information. If special processing is required the generate DIP, the function provides a processing request {internal flow} to the Process Data function itemizing the data object in the staging area which require processing and the processes to be applied. The processing function will provide a notice of completed processing {internal flow} and identify output data objects in the staging area.

* The **Process Data** function receives a processing request from the Generate DIP function. It accesses data objects in staging storage and applies requested processes. The types of operations which may be carried out include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or output formats, and other specialized processing (e.g. image processing). The Process Data function provides a notice of completed processing {internal flow} to the Generate DIP function upon completion of processing.

* The **Delivery** function handles both on-line and off-line deliveries of DIPs to consumers. For on-line delivery it accepts a DIP and prepares it for distribution in real time via Access **{4.1.7b}** over communication links. It identifies the intended recipient, the transmission procedures requested, and the DIP in the staging area to be transmitted. For off-line delivery it retrieves the DIP from the prepare DIP function, prepares packing lists, bills of lading and other shipping records, and then ships the **DIP {4.1.7f}**. When the DIP has been shipped, a **notice of processed order {4.1.7e}** is returned to Administration. This function also calculates and records **billing information {4.1.7i}** for delivered orders and supplies them to the accounting function in Administration.

* The **Monitor Requests** function will track a dissemination request from inception to receipt of data by the consumer. It sends an **order confirmation {4.1.7h}** to the consumer, and notifies a consumer when his order has been executed. The confirmation fully identifies the order including order number, date of order, date of execution, identity of data requested, and method of distribution. The Monitor Requests function tracks every order from inception to delivery confirmation. Operations personnel are able to query the pending order file to determine the number and content of each unfilled order. As each order is filled (either by automated or manual means), it is removed from the pending order file. The Monitor Requests function also has the ability to execute a standing order (a standard query and report procedure), based on elapsed time or some other trigger function.

4.1.8 DATA FLOW AND CONTEXT DIAGRAMS

The flow of data items among the OAIS functional entities is diagrammed in this section. **Figure 4-2** shows the more significant data flows. The flows associated with the Administration are generally support background activities of the other entities. To avoid complication of Figure 4-2, these background flows are illustrated in the context diagrams of **Figure 4-3** .

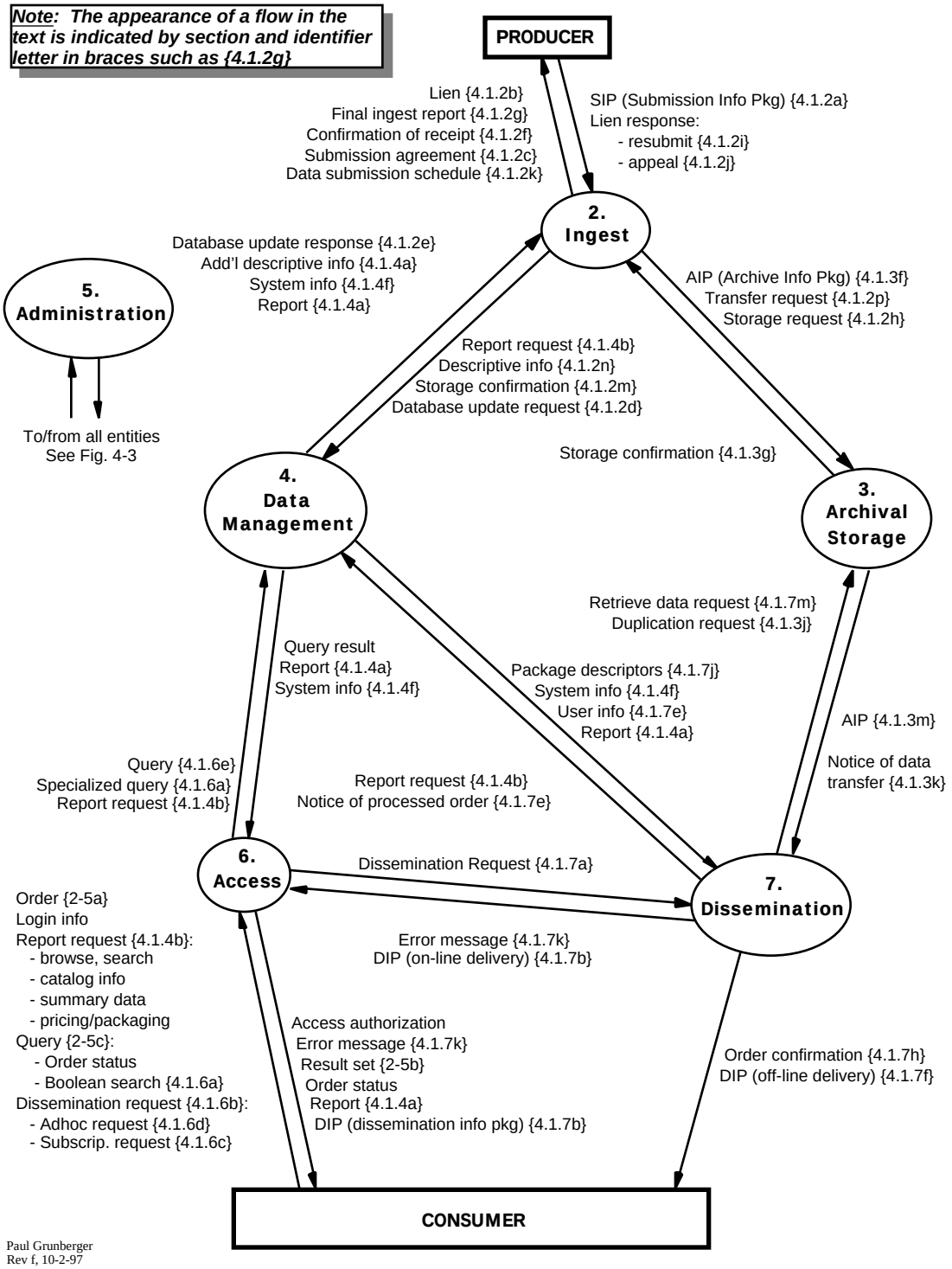


Figure 4-2. OAIS Data Flow Diagram

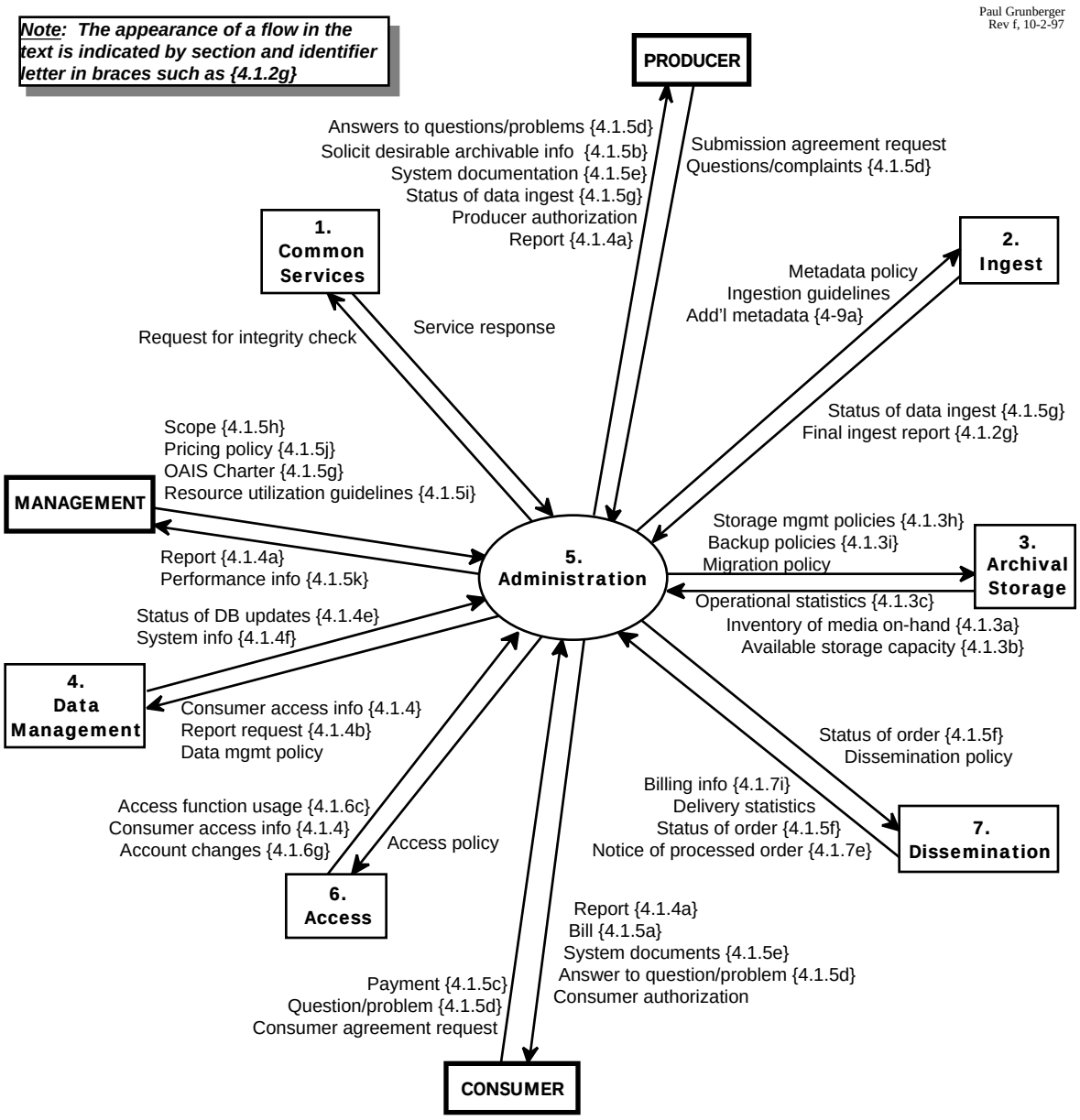


Figure 4-3. Administration Context

4.2 INFORMATION MODEL

This section builds on the concepts presented in section 2.2 to further describe the pieces of information that are exchanged and managed within the OAIS. This section also defines the specific information objects which are used within the OAIS to preserve and access the information entrusted to the archive. This more detailed model of OAIS related information structures is intended to aid the architect or designer of future OAIS systems. The structures discussed in this section are conceptual and should not be taken to imply any specific implementations.

As discussed in Section 2. the primary goal of an OAIS is to preserve information for a designated community over a indefinite period of time. In order to preserve this information an OAIS must store significantly more than the contents of the object it is expected to preserve. This section analyzes those classes to fully describe the object classes of data associated with an OAIS. This section uses **Object Modeling Technique** (OMT) [Annex D and Reference 2] diagrams to illustrate the concepts discussed in the text.

Section 4.2.1 expands on the OAIS Information discussions in section 2.2 to provide the reader with enough background concepts to understand the discussions in the later sections. This section also provides OMT diagrams to illustrate the Information Package variants concepts discussed in section 2.2.4 . Section 4.2.2 provides a more detailed view of the information required for effective long term preservation of information while Section 4.2.3 describes the conceptual objects and containers that represent the contents of an OAIS.

4.2.1. BACKGROUND CONCEPTS

Section 2.2 introduces the concept of information being a combination of data and Representation Information. This concept is illustrated by the OMT diagram 4-5. The **Information Object** (IO) is composed of a Data Object which is either physical or digital and the Representation Information that allows for the full interpretation of the data into meaningful information. There are many types of information that are needed for the long-term preservation of information in an OAIS and a taxonomy of these information types. A detailed discussion of Representation Information is presented in section 4.2.2.

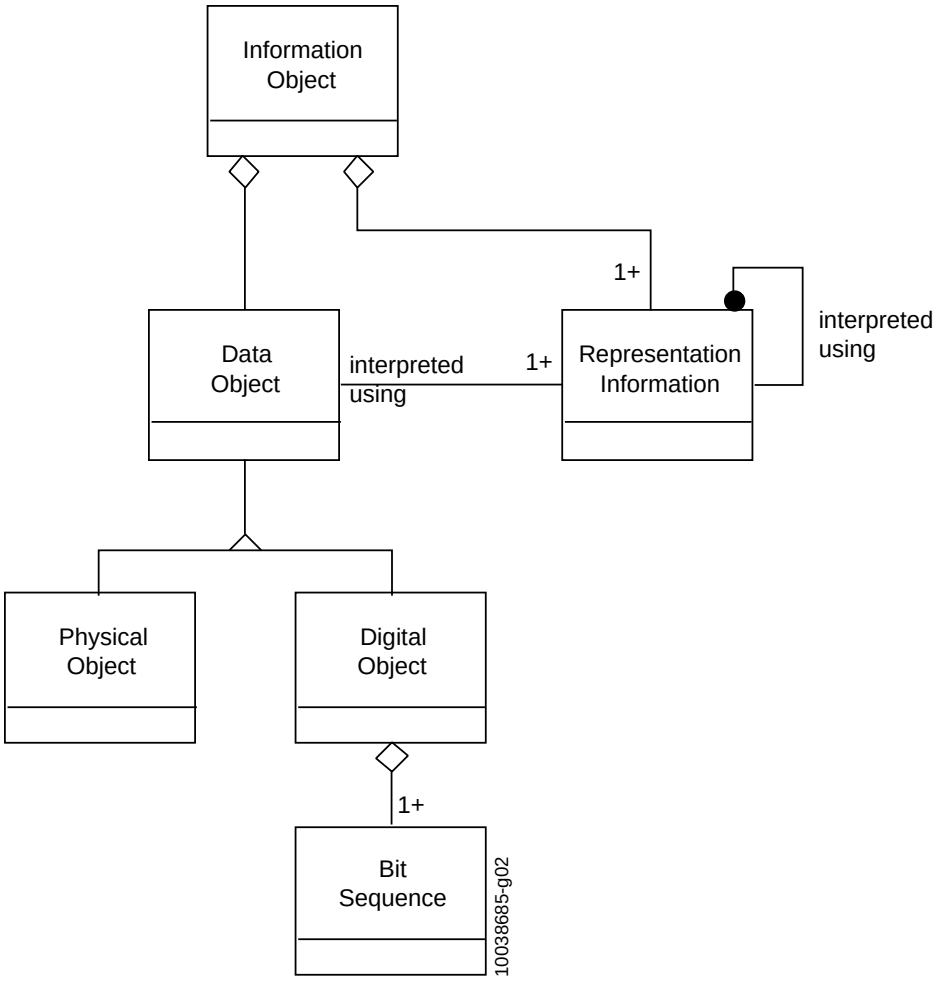


Figure 4-5. Information Object

The basic currency of the archival process is the Information Package (IP) . The OMT diagram in figure 4-6 illustrate the conceptual view of an IP. An information package is a container which contains two types of information objects, the Content Information (CI), the Preservation Description Information (PDI), and and is associated with Packaging Information (PI).

The CI is that information which is the primary target of preservation. It may be the content of a hardcopy document, or it may be the bit content of a CD-ROM file together with the Representation Information for those bits. Only after the CI has been clearly defined can an assessment of the PDI be made. The PDI is information about the CI and is needed to turn the CI into information that may be preserved over the long-term. Finally, the Packaging Information is that information which, either actually or logically, binds and relates the components of the package into an entity. For example, if the CI and PDI are identified as being the content of specific files on a CD-ROM, then the PI may be the ISO-9660 volume/file structure on the CD-ROM. The PI, in this case, may also include the physical CD-ROM disk. These choices are the subject of local archive definitions or conventions.

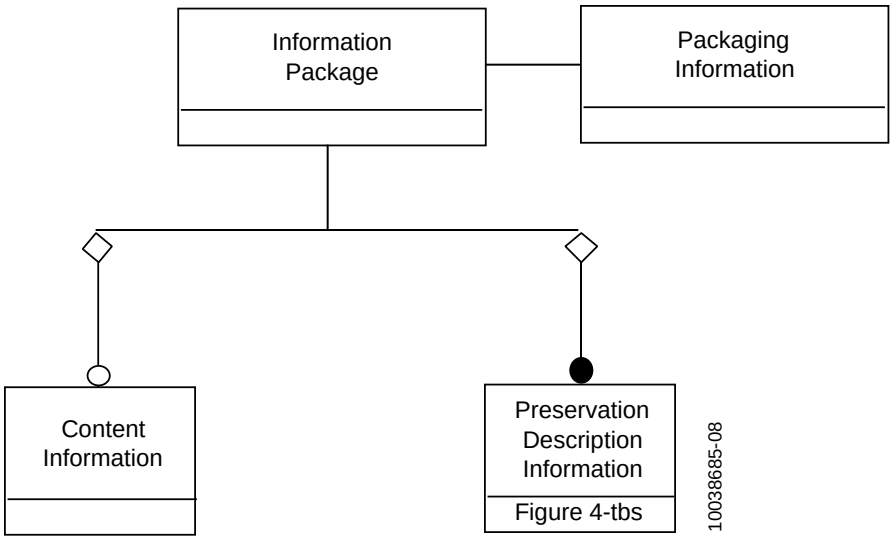


Figure 4-6. Information Package Contents

There are three subtypes of the IP defined in section 2.2, the Submission Information Package (SIP),the Archival Information Package (AIP), and the Dissemination Information Package (DIP). The definitions of these package types in section 2 is based on the function of the archival process which uses the package. This taxonomy of IP types is shown in Figure 4-7.

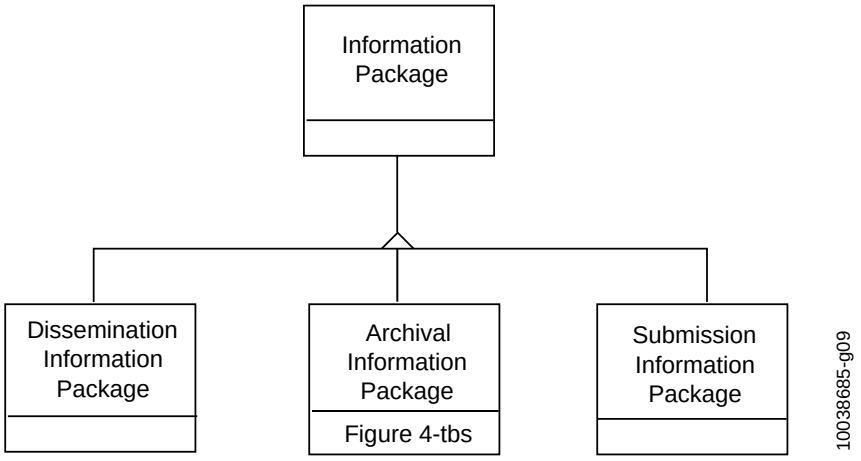


Figure 4-7. Information Package Taxonomy

There is no difference in the classes of Information Objects contained in the AIP, SIP or DIP. The difference is the optional versus required nature of specific information within the information package. The AIP must include a Content Information object, a complete set of Preservation Description Information. In addition the AIP must be associated with a set of

Packaging Information the provides delimitation of the AIP in the OAIS Archival Storage. The SIP and DIP may contain any of these IOs however, since it is not required that information be archived in the same form in which it is transferred there are no mandatory IO types within an SIP or DIP. There is mandatory associated Packaging Information which enables the consumer of the SIP or DIP to extract the digital information from the media.

Though in many cases the information submitted to an OAIS in one SIP is the complete information needed to create one AIP; there are often good reasons the content information of a SIP or DIP is different from the content information in the associated AIP. This means that there can be one-to-one, many-to-one, one-to-many, and many-to-many mappings between SIPs and AIPs. Here are some examples:

- *One SIP - One AIP* —A government agency is ready to archive its electronic records from the previous fiscal year. All of the year's records are placed onto magnetic tapes that are submitted as one SIP. The archive stores the tapes together as a single AIP.
- *Many SIPs - One AIP* — A satellite sensor makes observations of the Earth over a period of one year. Every week all of the latest sensor data are submitted to the archive as a SIP. The archive has a single AIP containing all of the sensor's observations and the latest SIP is merged into that AIP.
- *One SIP - Many AIPs* — A company submits financial records to an archive as one SIP. The archive chooses to store this information as two AIPs: one which contains public information and the other which contains sensitive information. This makes it easier for the archive to manage access to the information.
- *Many SIPs - Many AIPs* — An oil and gas company collects information on its wells. Every year it submits a SIP containing all of the well status information to an archive. The archive maintains one AIP for each oil or gas field and breaks out the information on each well to the proper AIP based upon its geographic coordinates.

4.2.2. LOGICAL MODEL FOR ARCHIVAL INFORMATION

This section develops an information model for the information that is preserved in an OAIS. This section builds on the discussions in section 2.2 and 4.2.1 about the types of supporting information needed to enable long term preservation and the role of Representation Information in converting data objects to information.

There are several types of Information Objects that are used in the OAIS. Figure 4-8 shows a taxonomy of information objects.. The objects are categorized by their function in an IP into content objects, preservation description objects and packaging information objects. The following sections discuss the contents of each of the types of Information Object .

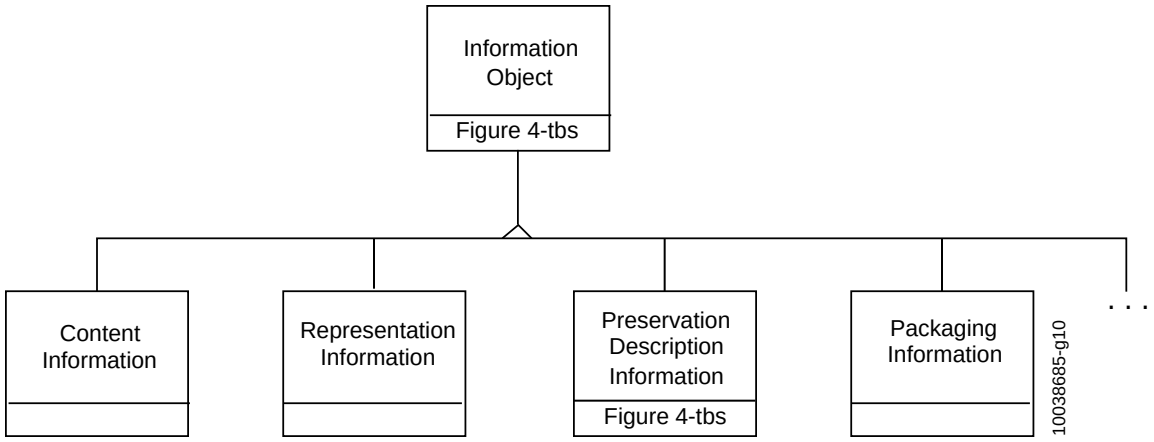


Figure 4-8. Information Object Taxonomy

4.2.2.1. Content Information

The **Content Information** is that information which is the primary object of preservation.. The Content Information can be viewed as a primary **Data Object** together with its Representation Information as shown in Figure 4-5 . The Data Object in the Content Information may be either a Digital Object or a Physical Object (e.g., a physical sample, microfilm).

4.2.2.2. Representation Information

The CI may be expressed as either a physical object (e.g., a moon rock) together with some **Representation Information**, or it may be expressed as a digital object (i.e., a sequence of bits) together with the Representation Information giving meaning to those bits.

The Representation Information accompanying a physical object like a moon rock may give additional meaning, as a result of some analysis, to the physically observable attributes of the rock. This information may have been developed over time and the results, if provided, would be part of the CI.

The Representation Information accompanying a digital object, or sequence of bits, is used to provide additional meaning. It typically maps the bits into commonly recognized data types such as characters, integers, and reals and into groups of these data types. It may associate these with higher level meanings which can have complex inter-relationships that are also described.

This section further addresses the Representation Information object when the Data Object is specialized as a Digital Object.

The Digital Object, as shown in Figure 2-2, is itself composed of one or more bit sequences. The purpose of the Representation Information object is to convert the bit sequences into more meaningful information. It does this by describing the format, or data structures, which

are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc. These common computer data types, and aggregations of these data types, are referred to as the Structure Layer information of the Representation Information object. These structures are commonly identified by name or by relative position within the associated bit sequences.

The Representation Information provided by the Structure Layer is seldom sufficient. Even in the case where the Digital Object is interpreted as a sequence of text characters, and described as such in the Structure Layer, the additional information as to which language was being expressed should be provided. This higher layer information is referred to as the Semantic Layer, although in reality each layer provides its own set of semantics. When dealing with scientific data, for example, the information in the Semantic Layer can be quite varied and complex. It will include special meanings associated with all the elements of the Structural Layer, and their inter-relationships. An expansion of the Representation Information object is given in Figure 4-9.

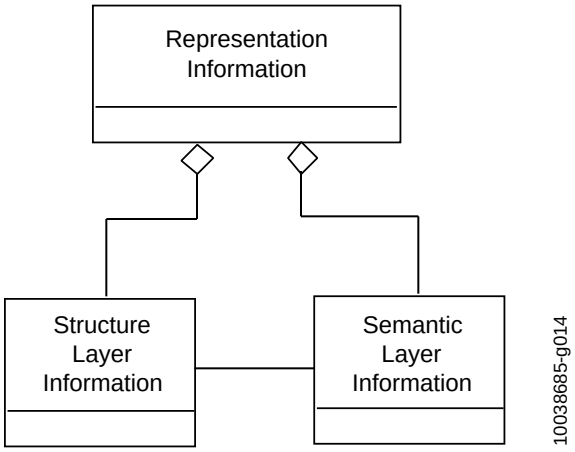


Figure 4-9. Representation Information Object

The Semantic Layer can also be viewed as the "Object" layer. Taking an object oriented view of the combination of the Representation Information and its Digital Object (i.e., Content Information object), queries applied to this object are addressed to the Semantic Layer. The object oriented methods translate these queries into actions on the Structure Layer elements, and ultimately to the bit sequences, with results reported back in Semantic Layer terms. Such software methods associated with a Content Information object provide useful services as long as the software executes properly. However for indefinite long-term information preservation, a full and understandable description of the Representation Information is essential. This can be a particular challenge for the preservation of scientific type data as there are few standards for how to express this type of information and archives need to ensure this information is understandable to the designated consumer communities.

All this additional Representation Information (both semantic and syntactic) is needed to

fully transform the bits of a file into the Content Information. In principal, this even extends to the inclusion of definitions (e.g., dictionary and grammar) of the natural language used (English in this example). Over long time periods the meaning of natural language expressions can evolve significantly in both general and in specific discipline usage. As an aside, it is clear from this example that, in general, 100% information preservation for the indefinite long-term is a goal that is not practically achievable.

Representation Information may be expressed in physical forms(e.g., a paper document) or in digital forms. When the Representation Information is in digital form, additional Representation Information is needed to understand the bits of the Representation Information. In principle, this recursion continues until physical forms are encountered. For example, Representation Information expressed in ASCII needs the additional Representation Information for ASCII, which may be a physical document giving the ASCII standard. Because each Representation can be composed of multiple components, each with is own Representation, the result can be described as a **Representation Net**. In practice, the recursive chain of the representation net is also broken when there is widely available software that understands a particular representation, such as ASCII display software. Nevertheless, this situation is dynamic as the technology and standards evolve and is a preservation issue needing continuing attention

As a practical matter, the OAIS needs to have enough Representation Information associated with the bits of the Data Object in the Content Information that it feels confident that it, and those expected to use the Content Information, can enter the Representation Net with enough knowledge to begin accurately interpreting the Representation Information. This is a significant risk area for an OAIS, particularly for those with a narrow discipline focus, because jargon and apparently widely understood terms may subsequently be found to be quite temporary.

As a complex example, consider an electronic file containing a sequence of values obtained from a sensor looking at the Earth's environment. There is a second file, encoded using ASCII, that provides information on how to understand the first file. It describes how to interpret the bits of the first file to obtain meaningful numbers, it describes what these numbers mean in terms of the physics of the observation being conducted, it gives the date and time period over which the observations were made, it gives an average value for the observed values, and it describes who made the observations. These two files are submitted to an OAIS for preservation.

Assume that the OAIS determines that the Content Information to be preserved is the observed bits together with their values as numbers and the physics meaning of these numbers. This information is conveyed by the bit sequence within the first file together with the **Representation Information** from the second file needed to transform the first file’s bits into meaningful physical values. Note that neither the first file’s underlying media nor the particular file system carrying the bits is part of the Content Information. From the second file only part of its content is considered a part of the Content Information and this is the part that enables the transformation of the bits from the first file into meaningful physical values. In fact this second file does not carry all the Representation Information needed to make this

transformation because the following additional information is needed:

– Information that the second file is encoded in ASCII so that it can be read as meaningful characters;

– Information on how the characters are used to express the transformations from bits to numbers to meaningful physics values. This information, typically referred to as a combination of format information and data dictionary information, may also include instrument calibration values and information on how the calibrations are to be applied. All this information may be widely understandable once the ASCII characters are visible because it has all been expressed in English (or some other natural language), or some of it may be in more structured forms that will need additional Representation Information to be understood.

Therefore the Representation Information of the second file needs additional Representation Information, and this information may need additional Representation Information, etc., forming a linked set of Representations of Representations. This is a good example of the complex Representation Net.

Recall that in the example above, there was a determination that the Content Information consisted of the observed sensor values and their meanings. This is by no means the only determination that could have been made. It could just as easily have been determined that the Digital Object of the desired Content Information was the bit sequences within the first file together with the all the bit sequences within the second file. The fact that some of these latter bit sequences are used to interpret the first files bit sequences is just an example of a set of bits that is somewhat self-describing. It is irrelevant that some of the bits in the second file are the basis for information on the date and time period over which the observations were made, the average value for the observed values, and who made the observations. Once it has been determined that all these bits constitute the Digital Object of the Content Information, then the Representation Information is that information needed to turn them into meaningful information. How extensive this meaning is to be carried and how far the Representation Net needs to be carried are local issues for the OAIS and its related producer and consumer communities.

The extent to which the CI is understandable to a designated community depends largely on the nature of the Representation Information. It needs to be written using constructs which are understandable to that community and it needs to be sufficiently complete to convey all the information intended. Narrowly drawn or specialized communities may need minimal Reference Information to understand a particular CI, but in such cases extra care needs to be exercised to ensure that the natural evolution of what is commonly understood in those communities does not effectively cause information loss from the CI.

4.2.2.3. Preservation Description Information

In addition to the content object the Information Package must include a set of IOs which will allow the understanding of the content objects over an indefinite period of time. The specific set of IOs which are required for this function are called collectively called

Preservation Description Information (PDI). The PDI IOs are a specialization of the IO shown in figure 4-6 in that the data object must be digital while the data object in an IO may be a digital object or a physical object. This restriction is due to the fact that the focus of this reference model is digital archives. PDI is that information which is necessary to adequately preserve the particular Content Information with which it is associated. It is specifically focused on describing the past and present states of the Content Information, ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered. This information is typical for all types of archives and has been classified in the context of traditional archives. However, the class definitions must be extended for digital archives. The following definitions are based on the categories discussed in the paper "Preserving Digital Information". Figure 4-10 in an OMT diagram which illustrates the taxonomy of PDI types.

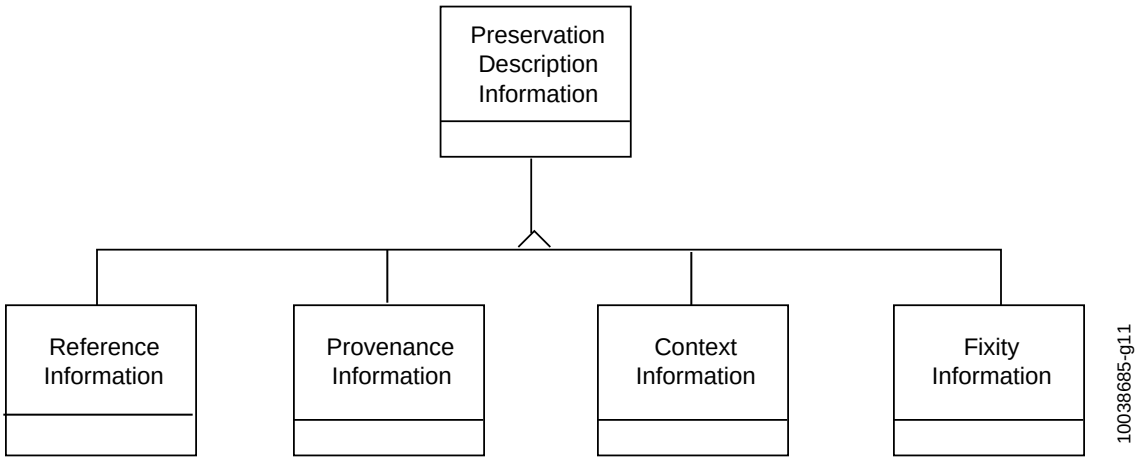


Figure 4-10. Preservation Description Information

– Provenance Information: This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This give future users some assurance as to the likely reliability of the Content Information. This information may be thought of as a special case of Context Information described below.

– Reference Information: This information identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Information.

– Context Information: This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere.

– Fixity Information: This information documents the authentication mechanisms, and it provides any authentication keys used to ensure that the particular Content Information object has not been altered in an undocumented manner.

The OAIS needs to explicitly decide what the Content Information is in order to be able to ensure that it also has the necessary PDI which is needed to preserve the Content Information. Deciding what is the Content Information, for a given set of information, may not be obvious and may need to be negotiated with the information producer. Once the Content Information has been determined, it is possible to assess the Preservation Description Information.

4.2.2.4. Packaging Information

Finally, the **Packaging Information** is that information which, either actually or logically, binds and relates the components of the package into a physical entity on a specific media. This packaging information consists of all the information necessary to delimit the IP and to locate the IP on the media. For example, if the CI and PDI are identified as being the content of specific files on a CD-ROM, then the PI may be the ISO-9660 volume/file structure on the CD-ROM. The PI, in this case, may also include the physical CD-ROM disk. These choices are the subject of local archive definitions or conventions. The Packaging Information does not necessarily need to be preserved by an OAIS since it does not contribute to the Content Information or the PDI, however, there are many cases where the archive is required to exactly reproduce the SIP as a DIP. In these cases the Packaging Information must be preserved and reproduced. The OAIS must also ensure that no PDI or Content Information is hidden in the naming conventions in directory or file structures.

4.2.3 LOGICAL MODEL OF INFORMATION IN AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

There are two major functions than an archive supplies to the community, long term preservation of information and effective access to the preserved information. This section discusses the conceptual information structures required to accomplish these functions.

Other than the replacement of the “digital objects” with the more structured and complex “archive information unit,” which enables the long-term preservation of information. The OAIS logical data model has been based on the Z39.50 Digital Collections Profile [reference 3], which is being widely used as a base standard for digital collection access and digital library access. . A more detailed view of the Z39.50 Digital Collections Profile and its relationship to the OAIS information model can be found in Annex C of this document.

4.2.3.1. The Archival Information Package

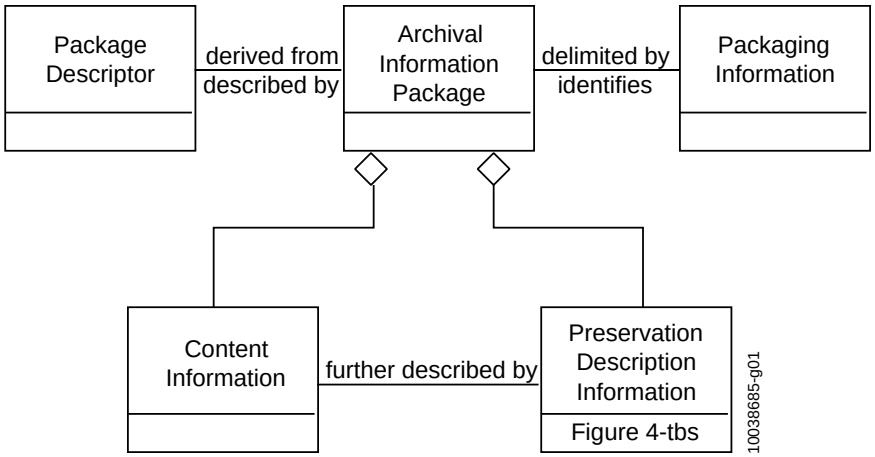


Figure 4-11. Archival Information Package(AIP)

An **Archival Information Package** (AIP) which is modeled in Figure 4-12 is a specialization of the IP which is defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite-long term, preservation of a designated Information Object. The AIP is itself an information object that is a container of other information objects. Within the AIP is the designated information object and it is called the Content Information.

Also within the AIP is an information object called the **Preservation Description Information (PDI)**. The PDI provides additional information about the Content Information and is needed to make the Content Information meaningful for the indefinite long-term.. An OAIS must clearly understand what constitutes the Content Information for each specific case in order to ensure that the corresponding Preservation Description Information fully performs its function.

The Preservation Description Information requirements in an AIP are must more stringent than the requirements for Preservation Description Information in an IP. While no PDI objects are mandatory in an IP, all PDI information must be present in an AIP. This is illustrated in Figure 4-12. In this case PDI is used as a container object which contains all the PDI object types.

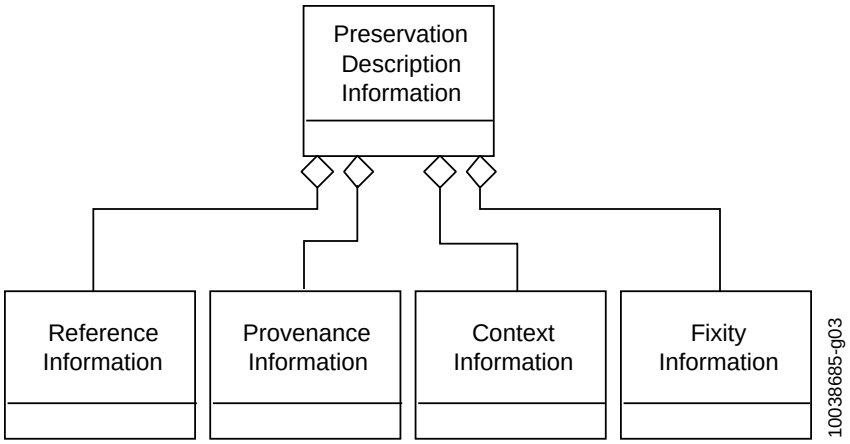


Figure 4-12. PDI in An Archival Information Package

The AIP is delimited and identified by the **Packaging Information**. The **Packaging Information** may actually be present as a structure in the AIP or may be virtual in that it is contained in the OAIS Archival Storage function. However, the delimitation and internal identification functions must be well defined in an OAIS.

The AIP objects described above provide the information necessary to enable the long term preservation function of the archive. In addition to preserving information, the OAIS must provide adequate features to allow consumers to locate information of potential interest, analyze that information, and order desired information. This is accomplished through the use of **Package Descriptor** which contain the data that serves as the input to documents or applications called Access Aidswhich can be used to locate, analyze or order information from the OAIS. The Package Descriptor is not required for the long-term preservation of the content information but is needed to provide visibility and access into the contents of an archive. The contents of the Package Descriptor are further defined in the next section.

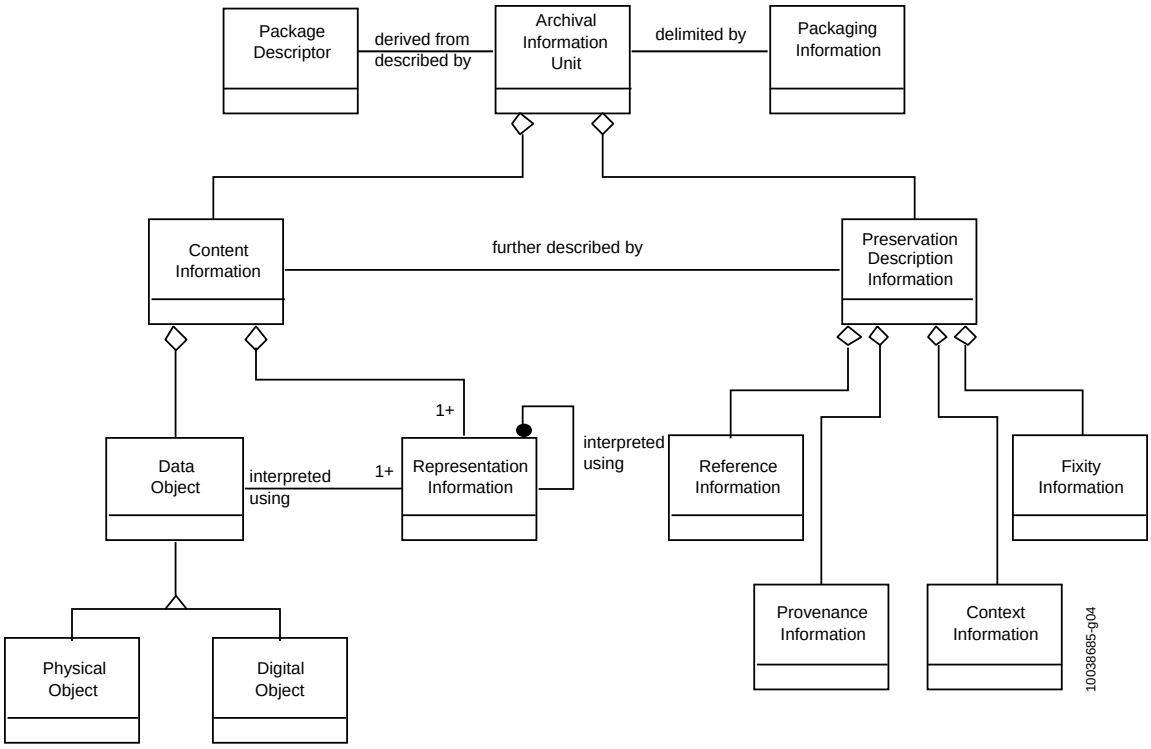


Figure 4 -13. Archival Information Unit (Detailed View)

Figure 4-13 gives a detailed view of the Archival Information Package by expanding the PDI and the Content Information. All the "contains " relationships discussed in this section are logical containment relationships. This type of containment relationship may be physical or may be accomplished via a pointer to another object in storage.

4.2.3.1. Specializations of the AIP

Two important specializations of the AIP are discussed in this section, the **Archival Information Unit (AIU)** and the **Archive Information Collection (AIC)**. Figure 4-14 is an OMT diagram illustrating this specialization. Both AIU and AIC are subtypes of the AIP and as such contain constructs to enable both long term preservation and consumer access. The AIU represents the primary type used for the preservation function. The AIC organizes AIPs along a thematic hierarchy which can support flexible and efficient access by the consumer community. The differences between AIUs and AICs is the complexity of their Content Information and their Package Descriptors.

To aid in the understanding of these constructs we will use an example of a company setting up an OAIS of digital versions of films. This example will focus on the information content of constructs in an AIP. Section 4.3 and the Illustrative Scenario in Section 7 will illustrate more of the details or the information transformations and dataflows in an OAIS.

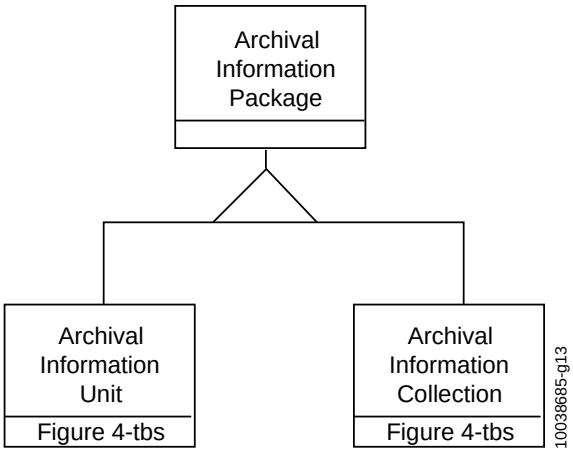


Fig 4-14. Archival specialization of the AIP

4.2.3.1.1. Archive Information Unit

The AIUs can be viewed as the "atoms" of information that the archive is tasked to store. A single AIU contains exactly one content information object (which may consist of multiple physical files) and exactly one set of PDI. When an SIP is ingested into the OAIS a **Unit Descriptor** , which is a subtype of a Package Descriptor is created by extracting information from the CI and the PDI and appending it to the unique ordering number and cost.

In the example of a digital film OAIS, the AIU for a single film can be viewed two objects, one containing a digital image of the film in a proprietary format (CI) and the other containing facts about the film such as date of creation , featured actors, director, producer, sequels, movie studio, and a checksum to ensure the integrity of the digital image (PDI). Since the OAIS reference model is implementation independent, these objects could be implemented as one file or multiple files. This type implementation dependent information is contained in the Packaging Information. When a movie is ingested into the OAIS a Unit Descriptor is created by extracting information from the CI and the PDI and appending it to the unique ordering number and cost.

The Unit Descriptor is a specialization of the Package Descriptor that always contains:

- a set of **Associated Descriptions** each of which describe the AIU content information from the point of view of a single **Access Aid**, and,
- a method for identifying and retrieving the AIU from Archival Storage

Figure 4-15 is an OMT diagram the illustrates the Unit Descriptor contents.

Access Aids are documents or applications that can be used to find the AIU, visualize the AIU or order AIU. The information needed for one access aid is called an Associated Description. A single Unit Descriptor may contain several associated descriptions depending on the number of different access aids that can locate, visualize, or order the associated AIU.

An important type of Access Aid is **Finding Aid** are applications that assist the consumer in locating information of interest. A single AIU may have a number of Associated Descriptions which describe the CI using different technologies. Additionally as new description extraction and display technologies become available an archive may want to update the Unit Descriptor associated with each of its AIUs to add a new Associated Description that utilizes the new technology to better describe the AIUs.

In the digital movie OAIS example, initially, there may be one Associated Description that is a free text description of a movie, another that is a five minute clip and another that is a row in a relational database that is used by film collectors to locate films of interest. After the archive has been operational for a period of time a technique for supplying compressed digital movies may be developed based on recording every tenth frame. The archivist may decide to create an additional type of Associated Description which is populated using the results of this new technique. If necessary, the user can run each of the AIUs contained in the archive though this compression technique and create a new Associated Description for each movie in his archiveor simply include this Asociated Description for new movies ingested into the OAIS.

Another important class of Associated Descriptions supply **Ordering Aids** allow the consumer to discover the cost of and order AIUs of interest. The Ordering Aids also allow users to specify transformations to be applied to the AIUs prior to dissemination. These transformations can include data object transformations such as subsetting, subsampling or format transformations. The transformations can also involve subsetting the PDI in an AIU prior to dissemination.

For example, the digital movie OAIS could allow a user to order a digital movie as a VHS tape, a laser disc or a JPEG object delivered on-line. Each of these would involve a format transformation and in theory an update to the PDI information in the AIP to create PDI for the DIP.

In addition to the Associated Descriptions, the Unit Descriptor also contains **Access Methods** which enable authorized users to retrieve the AIU described by the Unit Descriptor. In most current archives, only internal archive processes and operations personnel are authorized to use these Access Methods. However, as technology advances increase the processing power of the archive and the bandwidth between the archive and the user such access methods as “content based queries” are allowing the archive user direct read only access to the content objects of AIUs.

For example, a pattern recognition technique might be created for digital movies and the digital movie OAIS might offer a service to search its archives for large structures such as the pyramids or a New York skyline. Note: the sort of service is very processing intensive, if the results are generally useful, the archivist could summarize the results of this “content based query” into a new Associated Description. This technique is frequently referred to as data mining.

Although, the AIU and its associated Unit Descriptor provide all the information necessary for a consumer to locate and order AIUs of interest, it can be impossible for a consumer to sort through the millions of Unit Descriptors in a large archive. This problem is addressed in the OAIS through the Archive Information Collection discussed in the next section.

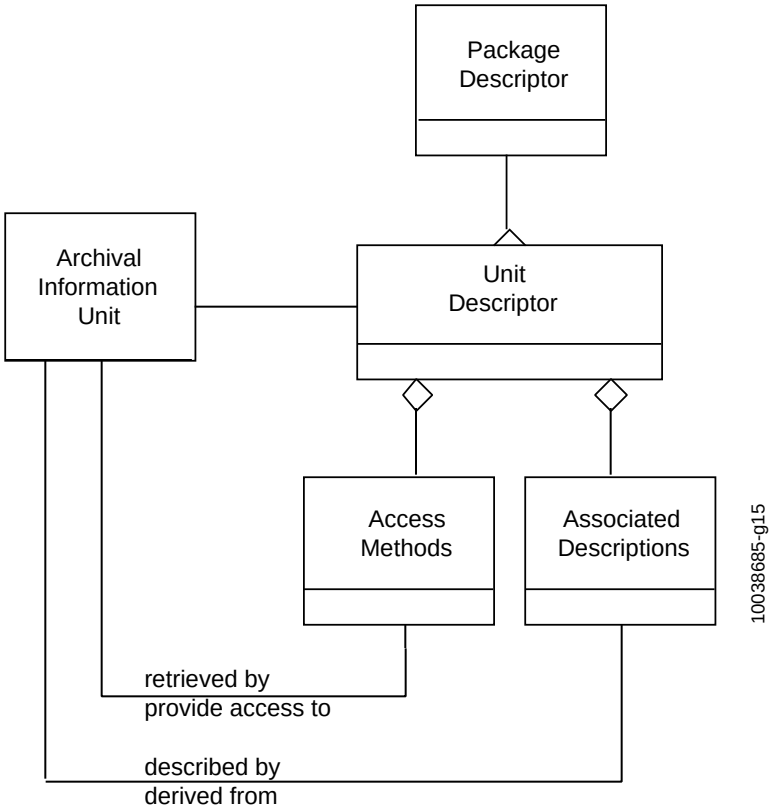


Figure 4-15-Unit Descriptor

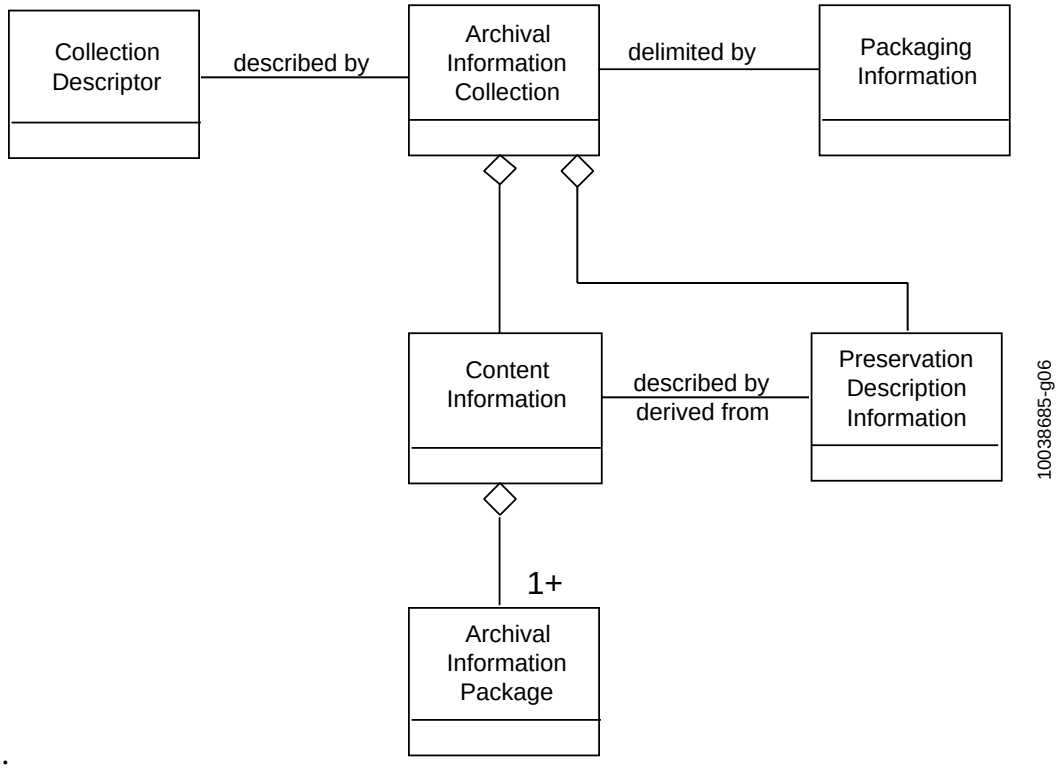
4.2.3.1.2 Archive Information Collections

The content information of an AIC is composed of complete AIPs each of which have their own CI, PDI, and associated PI and PD. These AIPs are then aggregated into Archive Information Collections (AIC) using criteria determined by the archivist. Generally AICs are based on the AIUs of interest having common themes or origins and a common set of Associate Descriptions. At a minimum all OAIS can be viewed has having at least one AIC which contains all the AIPs held by the OAIS.

A logical model of a AIC is shown in Figure 4-16. As in Figure 4-13 all the containment relationships are logical containment and may be physical or may be accomplished via a pointer to another object in storage. For example, the Content Information of an AIC can be created either by creating physical collections of the contained AIPs or by pointing to the contained AIPs. A single AIP can belong to any number of AICs.

For example the digital movies OAIS may have collections based on the subject area of the movie such as mystery, science fiction , or horror. In addition the archive may have AICs based on other factors such as director or lead actor.

An important feature of the AIC as shown in is the fact that an AIC is a complete AIP which contains PDI which provides further information about the AIC such as provenance on when and why it was created, context to related AICs, and fixity information. This is in addition to the PDI contained in member AIPs. This type of information is often necessary for a consumer to have confidence in the reliability of an AIC. In the digital movies OAIS example, the usefulness of an AIC of movies starring John Wayne is to some extent based on the provenance of when the collection was created or last updated.



.

Figure 4-16. Archive Information Collections Logical View

4.2.3.1.3 Collection Descriptors

The **Collection Descriptor** is a subtype of the Package Descriptor which has added structures to better handle the complex content information of an AIC. The Collection Descriptor, which is modeled in Figure 4-17 contains the information classes which are contained in the Unit Descriptor. The Access Methods of a Collection Descriptor provide a user with access to the entire Content Information of the associated AIC and the PDI for the AIC not for members of the AIC.

There are two types of Associated Descriptions in an Collection Descriptor:

- Associated Description that describe the collection as a whole (called Collection Description in Figure 4-17)
- Associated Description that separately describe each member of the collection (called Member Description in Figure 4-17)

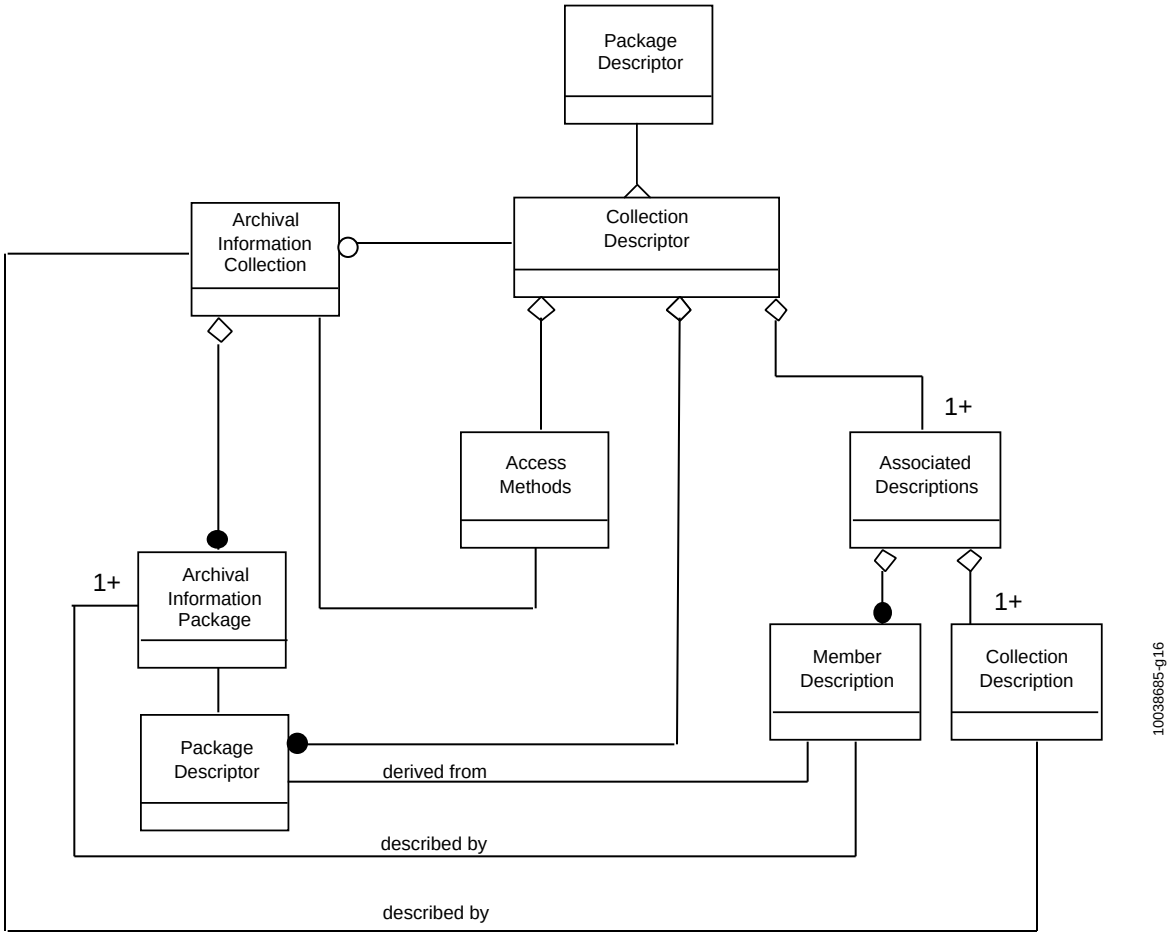


Figure 4-17. Collection Descriptors

In addition to the information contained in a Unit Descriptor, the Collection Descriptor optionally contains the Package Descriptors of the AIPs contained in the AIC. This containment relationship is logical in that the AIC mayeither the Package Descriptors of member IPs or, more commonly, pointers to the Package Descriptors of the member IPs. This list of the Package Descriptors for contained AIPs in an AIC could provide Access Aids a method to individual members of the AIC. It also allows an alternative concepts for the implementation of Member Descriptions . They could be implemented either in a centralized fashion as an Associated Description in the Collection Descriptor or in a distributed fashion by searching the Associated Description of each of member Package Descriptions.

Another important benefit of the Collection Descriptors is the ability to define new **Virtual**

Collections which are not organized around existing AICs. A Virtual Collection may be based on new data mining results or to reflect AIPs that observe common phenomena or areas of interest. To create a Virtual Collection, an archive would create a Collection Descriptor that did not have an associated AIC. The Collection Descriptor could have a customized Associated Member Description that documented the newly mined description data for each member AIP . A specialized finding aid could use this new Associated Member Description in conjunction with existing Associated Descriptions in the Package Descriptor information of each member AIP to locate AIPs of interest to the user. The Package Descriptors of contained AIPs would also supply an automated ordering aid with a method to order the IPs of interest to the consumer. Examples of Virtual Collection in a digital movies OAIS might be a new arrivals collection or a twenty most popular titles collection which is updated periodically.

Currently, Package Descriptors are stored in persistent storage such as database management systems to enable easy, flexible access and update to the contained Associated Descriptions. In addition to the Package Descriptors discussed in the previous sections, all the information needed for the operation of an archive would be stored in databases as persistent data classes. Some examples of this data are accounting data for customer billing and authorization, policy data, subscription data for repeating requests, and statistical data for generating reports to archive management. These classes are intended as examples rather than an exhaustive list of the data required for archive administration. Figure 4-18 ias an OMT diagram which illustrates the various types of "data management data" within the OAIS

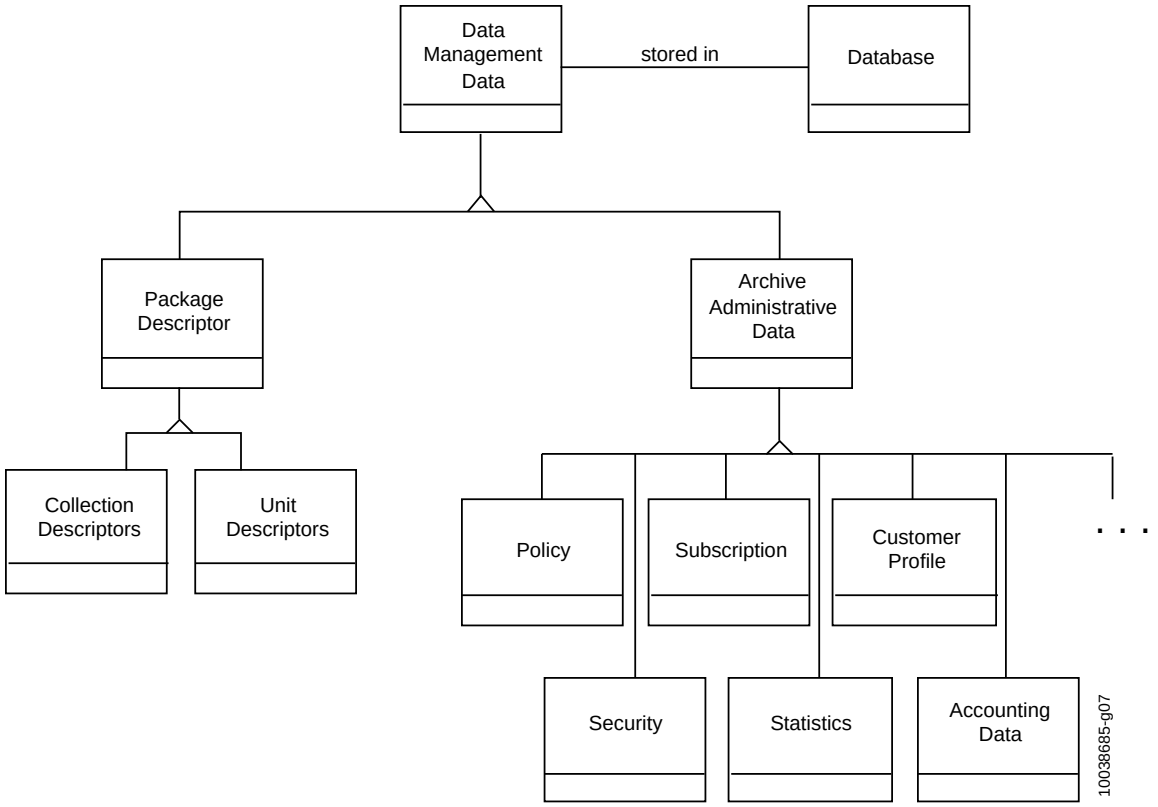


Figure 4-18. Data Management Data

4.3 HIGH LEVEL DATA FLOWS AND TRANSFORMATIONS

Figure 4-19 presents a high level data flow diagram which depicts the principle data flows involved in OAIS operations. These flows do not include administrative flows such as accounting and billing but concentrate on the flows between an OAIS and the other entities in its environment (producers and consumers) and internal OAIS flows that involve Information Packages.

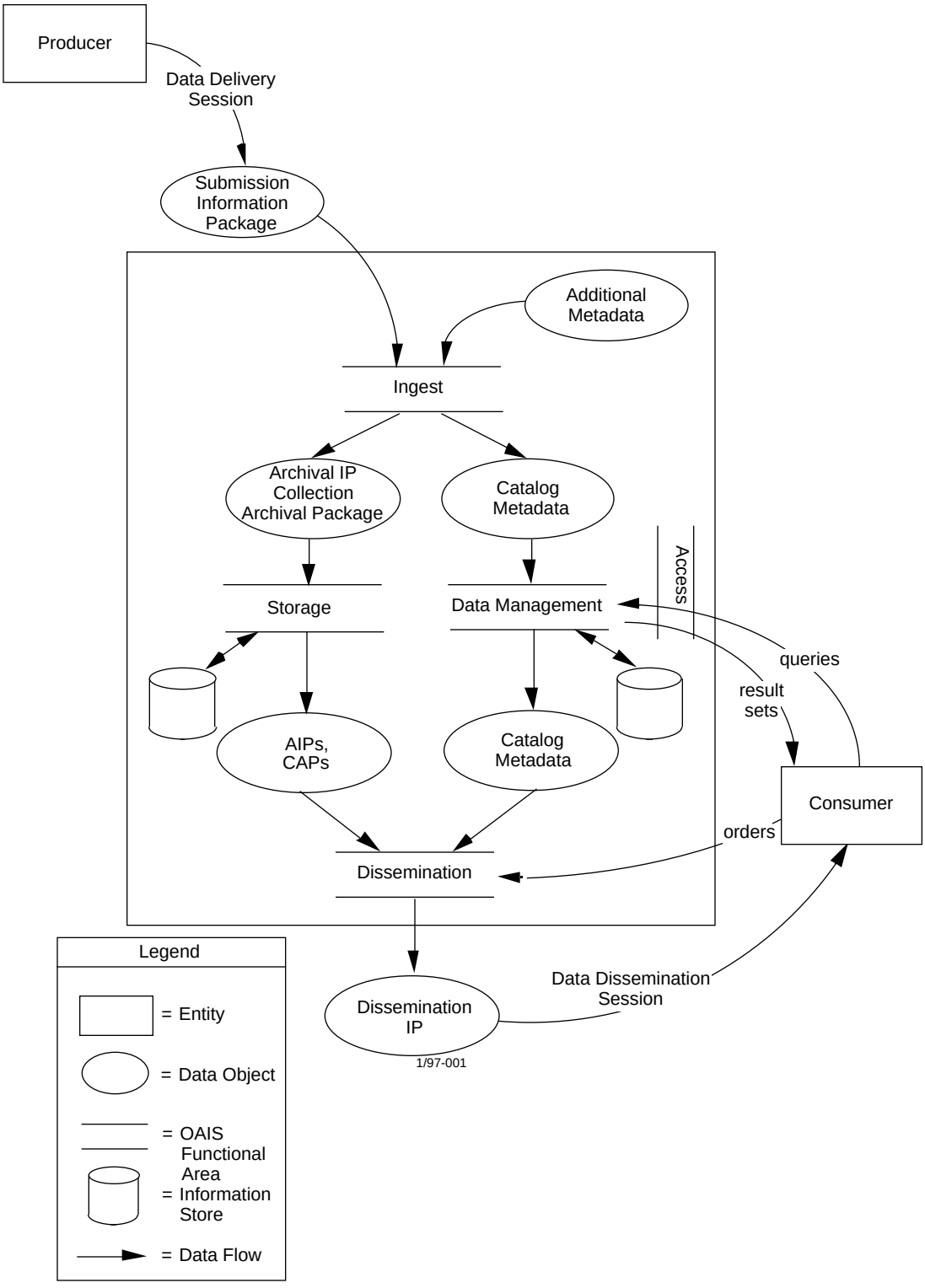


Figure 4-19. High Level Data Flows In an OAIS

When an information object is ingested by an OAIS, various metadata objects are required to assist in the long-term preservation of and access to the information contained in that object.

When the information object is stored or migrated, additional metadata objects are created to assist in the preservation and access process. The classes of these metadata objects are discussed in section 4.2 of this document and form the logical information model of the OAIS. This section discusses the range of potential transformations (both logical and physical) that may occur as an Information Package (IP) passes through the functional areas of the OAIS discussed in section 4.1 of this document.

It is important to note that at each point in these data flows, there is a complete Information Package consisting of Content Objects (e.g., images from a spacecraft instrument) and information describing those content object. There are three separate subtypes of Ips (i.e., the Submission Information Package, the Archive Information Package, and a Distribution Information Package) but the information provided in and with the original submission can be preserved in any of these packages.

4.3.1 DATA TRANSFORMATIONS IN THE PRODUCER ENTITY

The data within the data producer entity are private and may be in any format the producer desired. However, when the decision is made to store the data in an OAIS, the scientific investigators who are responsible for the data meet with archivists to negotiate a Submission Agreement.. This agreement spells out the content and format of the Submission Information Package(SIP). The SIP is an Information Package that is provided to the OAIS by the producer. The SIP consists of the scientific data plus the data that is necessary to assure that those data can be maintained by the OAIS and that the data can be interpreted and used by scientists who withdraw them from the OAIS at some time far in the future. These SIPs are periodically transferred to the OAIS in a Data Delivery Session. The number of data delivery sessions between an OAIS and a data producer can range from a single session in the transfer a final data product to multiple sessions a day in the case of active OAIS which store data for experiments which are still in process. The data delivery session can be logically viewed as sets of content data objects and description objects, although physically the description or metadata can be included in the digital objects (i.e., self describing objects) or divided into many separate descriptive items. In addition to the logical view of data (the SIP), the specification of a data delivery session must also include the mapping of the objects to the media on which they are delivered. This mapping includes the encoding of the object and description and the allocation of logical objects to files.

4.3.2. DATA TRANSFORMATIONS IN THE INGEST FUNCTIONAL AREA

Once the SIP are within the OAIS, their form and content may change. An OAIS is not always required to retain the information submitted to it in precisely the same format as in the SIP. Indeed, preserving the original information exactly as submitted may not be desirable. For example, the computer medium on which our images are recorded may become obsolete, and the images may need to be copied to a more modern medium. In addition, some types of information such as the Reference ID used to locate the Package within the OAIS will not be available to the producer and must be input after ingest to the OAIS.

The ingest process transforms the SIPs received in the data delivery session into a set of AIPs and Package Descriptors which can be stored and accepted by the Storage and Data Management functional entities. The complexity of this ingest process can vary greatly from OAIS to OAIS or from producer to producer within an OAIS. The simplest form of the process involves removing the Content Information, PDI and Package Descxriptors from the producer transfer media and queuing them for storage by the storage and data management functions. In more complex cases, the PDI and Package Descriptors may have to be extracted from the Content Information or input by OAIS personnel during the ingest function; the encoding of the information objects or their allocation to files may have to be changed; in the most extreme case, the granularity of the CI may be changed, and the OAIS must generate new PDI and Package Descriptors reflecting the newly generated information objects. In addition, the Ingest functional entity will classify incoming information objects and determine in what existing collection or collections each object belongs and will create messages to update the appropriate collections Descriptors after the AIPs are stored. During the ingest process, the OAIS must be highly aware not to unintentionally modify any of the information content in the Producer view. The OAIS is advised to save the producer media or copy the media into long-term storage as an ultimate reference if needed. It should be recognized that the saving of the producer media will not be permanent because of the issues discussed in Sections 2 and 4.2.2 of this document.

4.3.3. DATA TRANSFORMATIONS BY THE STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS

The storage and data management functional entities take the AIPs and Package Descriptors produced by the ingest process and merge it into the permanent archive holdings. The logical model of the ingested data should already be mapped into the logical model of the archives holdings, so the major transformation that occurs in this step is mapping the acquisition session from the ingest physical data model, which will tend to be on staging storage, to the permanent storage of the OAIS, which could range from database management systems (DBMSs) to hierarchical file management systems (HFMS), or any mixture of the above. The internal view of the OAIS is the permanent representation of the archived data, so all encodings and mappings must be well documented and understood. The process of transferring the ingest objects is frequently by a software process such as an HFMS driver or a DBMS. In this case, it is the responsibility of the OAIS to maintain an active copy of the software or careful documentation of the internal formats so the data can be transferred to other systems in the future without loss of information.

4.3.4. DATA FLOWS AND TRANSFORMATIONS IN THE ACCESS FUNCTIONAL AREA

When a Consumer wishes to use the data within the OAIS, he may use a finding aid to locate information of interest. These finding aids present data consumers with the logical view of the OAIS so the consumers can decide which information objects they wish to acquire. At a minimum, the access view is the high-level logical view of collections described in section 4.2.3. In most cases, the OAIS will have spent significant time and effort developing associated description and finding aids such as catalogs that will aid the user in locating

information objects or collections of interest. The consumer will establish a Search Session with the Access entity. During this Search Session, a Consumer will use the OAIS finding aids to identify and investigate potential holdings of interest. This searching process tends to be iterative, with a user first identifying broad criteria and then refining the criteria on the basis of previous search results. When the user has identified candidate objects of interest he may use more sophisticated visualization aids such as browse image viewers or animation to further refine his result set.

Once the Consumer identifies the OAIS holdings he wishes to acquire, he must issue an order request to the OAIS to acquire the data. This order can also specify any transformations the Consumer wishes applied to the AIUs in creating the Dissemination Information Package (DIP). The request triggers the dissemination process, which is discussed in section 4.1.7.

OAISs and external organizations may provide additional Associated Descriptions and finding aids that allow alternative access paths to the information objects of interest. As data mining technologies become more mature, it is likely that researchers will develop new and fundamentally different access patterns to information objects. It is important that an OAIS's access and internal data models are sufficiently flexible to incorporate these new descriptions so the general user community can benefit from the research efforts. A good example of this type of new associated description are a phenomenology database in Earth Observation, which allows users to obtain data for a desired event such as a hurricane, or volcano eruption from many instruments with a single query. It is important to note that such finding aids may become obsolete unless the data they require are preserved as parts of the AIPs they access.

4.3.5. DATA FLOWS AND TRANSFORMATIONS IN THE DISSEMINATION PROCESS

The dissemination process serves the data consumer in roles very similar to the way the ingest process serves the data producer. Through interactions with the access functional area, the data consumer produces a logical view of the desired information objects and descriptions to be include included in the Dissemination Information Package. At this point, the consumer issues an order request for this DIP that triggers the dissemination process, which negotiates a request agreement with the customer in which the physical details of the Data Dissemination Session such as media type and object format are specified.

The Dissemination functional area then contacts the Storage and Data Management functional areas and requests the AIPs and associated Package Descriptors necessary to populate the DIP requested by the consumer. The Storage and Data Management functional areas create copies of the requested objects in staging storage.

The Dissemination process transforms this set of the AIPs and associated Package Descriptors into a DIP and stores that DIP onto physical distribution (either physical or communications) media to be delivered to the data consumer in a Data Dissemination Session. The complexity of this transformation process can differ greatly on the basis of the level of processing services offered by the OAIS and requested by the data consumer in his order. In the simplest case, the DIP contains duplicates of the the AIPs and associated

Package Descriptors of interest from storage and data management function. In more complex cases, the desired CI may have to be extracted from the information objects or inserted into self-describing information objects, and the encoding of the information objects or their allocation to physical files may have to be changed. In the most extreme case, when the OAIS supports subsetting services, the granularity of the of the information objects may be changed, and the Dissemination process may generate AIPs and associated Package Descriptors reflecting the newly granularity..

5 MIGRATION PERSPECTIVES

The fast-changing nature of the computer industry and the ephemeral nature of electronic data storage media are at odds with the key purpose of an archive: to preserve information over a long period of time. No matter how well an archive maintains its current holdings, it is likely that over time much of the stored information will need to migrate to different media — today's data storage media can typically be kept at most a few decades before the probability of irreversible loss of data becomes too high to ignore — or to a different hardware or software environment to remain accessible. An OAIS should have a general plan for how each type of media is to be stored, monitored, and ultimately replaced. Media manufacturers and industry associations can provide much of the information needed to formulate this policy. An OAIS should also monitor all the Representation Information types associated with the Information Packages it handles and it should have a plan on how it will replace the Representation Information objects used when they are no longer effective (e.g., no longer supported by readily available software). The archive should also have general procedures to follow when updating hardware or software, and procedures to be used after migration to assure that information is preserved. Specific requirements for, or limitations on, the migration of individual Information Packages and their contents should also be spelled out in their associated Submission Agreements.

Three levels of archival information migration can be identified:

- Replication — A straight copy of data stored on a specific medium onto the same type of medium. This is typically done to extend the life of data beyond the storage life of the medium on which it resides. The physical and logical representations of the original Information Package are preserved and hardware and software that could previously access the package and its contents should still be able to do so.
- Repackaging — A change of physical or logical packaging (IP Packaging Information) that does not affect the package content (i.e., Content Information (CI) and Preservation Description Information (PDI)). Physical repackaging is the transfer of data from one medium to a different medium — typically to extend the life of the data beyond the technological life of a storage medium or to make the archive more compact by transferring data to a higher-density medium. Logical repackaging is a change to some or all of the logical context (volume, directory, and file organization, for example) in which the content to be preserved is embedded, without altering this content itself. Physical and logical repackaging often occur together. An example would be copying several CD-ROM disks to a higher-density Digital Video Disk, making the root directory of each CD-ROM a subdirectory under the root of the DVD-ROM. This changes the physical package — the data can no longer be read with a CD-ROM reader — and also the logical packaging since modifications to file pathnames are required, but it leaves unaltered the essential CI and PDI that was transferred from CD-ROM. (This assumes that the CD-ROM volume/directory bits were not a part of the original CI definition.)
- Transmutation — Migration where the representation of the Content Information or Preservation Description Information is modified. An example would be applying a loss-less compression scheme to the Content Information (but not the Packaging Information) to compress Content Information previously stored in the archive. (Note, if such

compression were applied to the whole IP at one time, including its content, it would only be a case of Repackaging because the decompression to access the package content would return the content unaltered.)

The general rule for archival information migration is that 'meaning' should be preserved, even though the representation of that information may change. Copied information is "equivalent" to the original if there is a known transformation that can generate the original information from the copy. Usually this means that information should not be deleted during migration unless it is redundant. Data may be modified or added as long as meaning is preserved. For example, ASCII character-coded data may be migrated to a new character set if the new code is a superset of ASCII (this is an example of transmutation). Each level of migration raises unique issues regarding the equivalence test, as discussed below. In practice, full equivalence may or may not be necessary or even achievable. In cases where full equivalence is not maintained during migration, then appropriate PDI additions should be included with the affected Information Package to describe how the copy differs from the original.

5.1 REPLICATION

The goal of replication is to retain a strict equivalence, including the physical form of the data. Thus a replica is equivalent to its original if both can be read with the same hardware and software and the original and copied Information Package match bit for bit after a medium has been replaced.

5.2 REPACKAGING

For simple physical repackaging, where the Packaging Information is copied onto a different type of medium without any other reformatting, the original and the copy can be said to be equivalent if the original Packaging Information bit sequence can be reproduced from the copy. Logical repackaging complicates the picture somewhat, dividing the issue of equivalence into two parts: preservation of Information Package content; and congruity of logical packages. Information Package content is preserved during migration if the original bit sequence of the content can be reproduced from the copy's Information Package content. Repackaged Packaging Information can be said to be congruent to the original if there is a transform that maps from the logical packaging of the copy back to the logical packaging of the original. This means, for example, that directory paths and file names can change, provided that there is a one-to-one mapping between the old and new names. This mapping could become part of the IP's Provenance Information. Congruence of packaging is not strictly required, but it helps to demonstrate that CI and PDI has not been lost or scrambled during migration and it can provide guidance when adapting software to support repackaged information. New information can be introduced during repackaging, but should typically be kept separate from the original material (for example, new material added as new files rather than modifying original files).

A significant problem that an OAIS can have is the handling of CI and PDI where the data Producer has embedded pointers to Packaging Information within the CI or PDI. A major

objective for the OAIS is to be able to do all Repackaging within the Archival Storage entity in a way that is transparent to services and functions outside the entity and does not affect the CI and PDI. This means that CI and PDI should only use pointers to Information Objects which are citable references or are globally unique identifiers. Accompanying the CI and PDI, as part of the Packaging Information, should be one or more standard mapping objects which map from the CI and PDI references/pointers to local Packaging Information identifiers (e.g., local path names).

Another problem is the occasional need to reproduce, in the Dissemination Information Package, a specific Packaging Information view that is not a standard for the OAIS internal operations. An example is the need to provide specific file names or the need to support a specific directory/file implementation so that, for an interim period, specific application software can access the DIP content without alteration. In such cases the OAIS will need to retain enough information in the Packaging Information so as to be able to reproduce the desired output forms.

5.3 TRANSMUTATION

The information conveyed by a specific data object, such as that supporting a particular Content Information object, is preserved if the original object can be precisely and completely reconstructed from its replacement. This allows for superficial changes to a data object -- for example, a change in byte ordering -- but also allows for more complex transformations.

For primitive data objects, a general set of rules can be given:

- For character-coded data, the character set may change as long as the new set can represent all of the old characters. For example, migrating ACII character data to the new Unicode standard would be admissible.

- For numeric data, two factors come into play: magnitude and precision. For integer numeric values, only the magnitude is important; integer numeric data can be changed to a different integer representation if the new format can represent the old data's maximum positive and negative values. For fixed-point real numbers, both magnitude and precision requirements must be addressed. Floating-point real numbers present the most difficult case, since both the magnitude of the exponent and mantissa components must be considered. More importantly, precision may change under different floating number formats. It should be noted that the common solution of storing real numbers as character-coded values is not guaranteed to preserve information since the conversions from binary format to character format and back may differ across hardware/software environments. One solution is to always use hardware that adheres to a standard — like the IEEE-488 standard — for the representation of numeric data.

For complex data objects — composed of a set of primitive objects — the equivalence principle allows for many kinds of updates, including changes in underlying representation

(using floating point rather than integer numbers), change in storage order (column major versus row major), and even lossless compression of the data. Some migrations, however, may not preserve information in the strictest sense. Examples are lossy data compression and remapping of data to a different map grid. The issue of information preservation then becomes a subjective matter to be decided by the information producers, archivists, and information users: if modifying the data does not alter significantly the results that users arrive at when they examine or employ those data, then the meaningful information has been effectively preserved. If there is doubt about whether or not significant information will be lost during a migration, it is recommended that the original information be disseminated out of the archive and a new information package generated, ingested back into the archive, and then maintained along with the original information.

5.4 UPDATING THE PRESERVATION DESCRIPTION INFORMATION

Whenever archived information is migrated, the Preservation Description Information associated with the affected Information Packages may need updating. Pure Repackaging should not require update to the PDI, but they should be tracked by the OAIS in any event. When transmutation is performed, representation information will typically also need to be updated to properly reflect the modified information content and this will require updates to the PDI.

6 ARCHIVE CLASSIFICATIONS

The Reference Model for an Open Archival Information System is meant to cover a wide range of possible implementations and therefore a variety of archives are expected to use the standards which will be based on the reference model. Given such a diversity, it is useful to have a uniform manner for describing or 'classifying' archives; both as a means of comparing two archive systems and of describing one's archive to one's management. While the state of describing archives has not yet progressed sufficiently far to rate them on an absolute scale, the points enumerated and discussed below do allow a context for such discussions.

1. Acknowledged Degree of Permanence

Classification criteria:

a) Temporary archive: archive has a defined lifetime and is not expected to exist beyond that time.

b) Permanent archive: archive is defined as permanent, or the date at which it may be terminated may be extended after additional review by management.

2. Digital Information Preservation Level

Classification criteria:

a) Bit Preserving: archive is responsible for preserving collections of bits

b) Information Preserving: archive is responsible for preserving bits as well as what those bits mean.

(This does not address the physical media/samples issue)

3. Degree of Opaqueness of AIP

The traditional view of archives is that the objects are opaque to the access function of the archive (ref. Digital Collections Profile). In other words, the access function can only process queries based on metadata (catalog entries) which have been explicitly stored for each object. This does not mean that information could not have been extracted from the objects as they were ingested (or at a later time) and that this information was attached to the object as metadata. This traditional barrier is beginning to breakdown in some archives (or digital libraries) and "data-mining" techniques may be available in some cases.

Classification criteria:

a) Objects are completely opaque to the access function.

b) Some information within an object is directly available to the access function.

c) Objects are available to the access function.

4. Dissemination Methods

Traditionally, archives have delivered ordered products on media (paper ,magnetic tapes, CD-ROM, FTP sessions). With the growth of communications bandwidth and Internet technologies there is a trend towards making archive access and dissemination seem like a single interactive process. This view can change the underlying architecture of the archive and severely restrict the maximum size of DIPs.

Classification criteria:

a) Non-electronically readable media (such as paper and microfilm)

b) Electronically readable media

c) Electronic transmission (such as Internet and modem)

5. Active vs. Final Archive

The traditional archive provides storage, access and dissemination of unchanging artifacts. However, many current scientific "archives" combine the functionality of data processing, data access and data dissemination for science products that are being created for the first time. These "active archives" must deal with all the issues of traditional archives but have several unique problems such as the ingest of new "versions" of currently archived objects, changing metadata for existing AIPs, and continuous ingest processing demands. Also the relative priority of the ingest, access, preservation and dissemination functions may differ in active archives vs. traditional archives.

Classification criteria:

a) Active archive

b) Final archive

6. Diversity of Collection

Archives may support any level of heterogeneity of the subject matter in their collections. The level of heterogeneity of the collections will affect several factors including:

- Degree of quality control
- Degree of user support
- Who oversee quality control
- Degree of discipline expertise (or specific data expertise)
- Number of SIP formats accepted
- Sophistication of finding aids

Classification criteria:

a) Project

b) Discipline

c) General

7. Institutional vs. Non-Institutional Archive

Ultimately, an archive serves the interests of the organization which gave it its charter and (one hopes) provides in some manner for its funding. However, the source of objects and/or the reason for storing the objects may or may not be directly related back to this organization. Often the community supplying or retrieving the objects is larger than the chartering organization or the chartering organization is only a part or representative of the community. If the archive is an integral part of the organization used primarily for preserving its own records, than it would best be described as an Institutional Archive.

Examples of Institutional archives: NARA, State government, Coca-Cola.

Classification criteria:

a) Institutional

b) Non-institutional

8. Archival Storage Types

Classification criteria:

a) Physical: includes physical samples, hard copy, film, etc.

b) Digital: Information archived is in digital forms

c) Both: Some information archived is in digital and some is in physical forms

9. Distributed vs. Centralized

This archive reference model does not specify whether the functional areas are physically centralized or distributed. However the current model assumes a single administration and management function which sets the operational policies for all functional areas. The use of the OAIS model with federated archives will drive additional administration and data flow considerations.

Classification criteria:

a) Centralized

b) Physically distributed, centralized administration/management

c) Physically distributed, federated administration/management

7 ILLUSTRATIVE SCENARIOS

The scenario in this section describes the flow of information into and out of a hypothetical archive and describes the organization and maintenance of the information within the archive. The scenario is based upon a specific example — an archive containing the medical records of a hospital — but the same principles found here apply to a wide variety of applications, including: archival of the records of a government agency; or the financial files of a multinational corporation; or the computer-aided design drawings for a company's products; or scientific data on a particular topic gathered by researchers from around the world. For all these applications the intent is to preserve information for an extended period — decades at least, even centuries in some cases.

We assume that the archive is already established and operational, with hardware, software, policies, procedures and personnel in place. We begin therefore with the preparations for placing a new kind of information into the archive. The hospital decides to archive the diagnostic images from a Computer Tomography scanner. The producers — members of the hospital's radiology department — meet with the archivists to pave the way for this submission. The archivists are not experts in CT imagery, so during this discussion the radiologists describe the data they will be submitting and how they expect the data will be used in the future. Together the producers and archivists draw up a *Submission Agreement* that spells out the form in which the data will be submitted, the frequency with which it will be submitted, the length of time for which the data are to be archived, the disposition of the data when that time expires, limitations on who may access the archived data, and so on. The submission agreement defines the content and format of the *Submission Information Packages* that will be provided to the archive. The SIP will include any or all of the following: (1) content information; (2) preservation information; and (3) packaging information.

Content information consists of *data objects* and *representation information.* Data objects are the most important part of the submission and the other parts of the SIP are provided chiefly to ensure the proper archival and eventual retrieval of the data objects. Data objects can be either *physical object* s or *digital objects*. An example of a physical object would be an X-ray image recorded on film; a digital object would be a 3-dimensional CT image recorded on a computer-readable medium. *Representation information* is the next most important part of the submission: it is information that is essential to understanding a data object but that is not provided as part of the data object itself, and which will likely not be available to users through outside sources. For either an X-ray or CT scan this might include some information on the patient — name, sex, age, and so forth — as well as information on the X-ray or CT device — brand and model, exposure settings, etc. A digital object like a CT scan will furthermore require some representation information that is not needed for a physical object: namely, a description of how the bits and bytes that make up the data object are arranged. For example, this kind of representation information would specify that the CT images are composed of pixels, each of which is an 8-bit unsigned integer number; that pixels are arranged in a certain order into image planes of a specific size, and that a specific number of image planes (in a certain order) form the 3-dimensional CT scan.

Preservation Information is additional information required to maintain the content information within the archive and to ensure that the content information can be interpreted and used by consumers at some time far in the future. Preservation Information comprises four kinds of information: Context; Provenance; Reference; and Fixity. Any or all of these may be included in a SIP. They may also be added by archivists when the SIP is ingested. Here are some examples of these kinds of information:

- *Context* — Information linking content information to a larger environment. For the images from our CT scanner, context information might include the ID number assigned to the patient by the hospital, which can be used to link the images to the patient's other records. Context information might also include a reference to a technical description of the CT scanner that might be needed by future doctors to properly interpret the images.
- *Provenance* —Information describing how the content information was processed or handled before being inserted into the archive. Take, for example, the data from a scientific instrument. The data submitted for archival have been calibrated to convert the unique output of the instrument into a physical quantity (temperature, pressure, etc.) that can be compared with other scientific data. The calibration that was used in this process would be included in the submission as provenance information.
- *Reference* —A reference is an identifier by which the content information is known outside of the archive. A simple example would be the ISBN number for a book stored in an archive in electronic form.
- *Fixity* — Information needed to validate or authenticate the content information. This could be something as simple as a checksum that verifies that the bits within a CT image have not been corrupted since the image was produced. For sensitive data, this could include information needed to encrypt and decrypt the content information.

When a SIP is ingested into an archive, it is stored as one or more *Archive Information Units.* An example of an AIU might be a CT scan. Frequently AIUs are aggregated within the archive into *Archive Information Collections*. An example would be all CT scans for a single patient. And AICs can be aggregated into larger AICs: for example, all CT scans of all patients seen during a given year. As part of the submission planning process, archivists determine how SIPs will map into AIUs and AICs. There can be one-to-one, many-to-one, one-to-many, and many-to-many mappings between SIPs and AICs. Here are some examples:

- *One SIP - One AIC* —A government agency is ready to archive its electronic records from the previous fiscal year. All of the year's records are placed onto magnetic tapes that are submitted as one SIP. The archive stores the tapes together as a single AIC, with each tape identified as an AIU.
- *Many SIPs - One AIC* — A satellite sensor makes observations of the Earth over a period of one year. Every week all of the latest sensor data are submitted to the archive as a SIP. The archive has a single AIC containing all of the sensor's observations and the latest SIP is merged into that AIC.
- *One SIP - Many AICs* — A company submits financial records to an archive as one SIP. The archive chooses to store this information as two AICs: one which contains public information and the other which contains sensitive information. This makes it easier for the archive to manage access to the information.

- *Many SIPs - Many AICs* — An oil and gas company collects information on its wells. Every year it submits a SIP containing all of the well status information to an archive. The archive maintains one AIC for each oil or gas field and breaks out the information on each well to the proper AIC based upon its geographic coordinates.

A SIP will typically consist of a set of files transferred to the archive either over a network or on storage media. If the submission is on storage media, the archive may simply store the submitted media or copy the information to different media before placing the information under the control of the *Storage Entity.* When a SIP arrives at the archive via a network transfer, the information is recorded onto appropriate media and maintained thereafter by the Storage Entity. Over time, the media on which information is archived will degrade and may become obsolete, and the information will need to be copied to ensure preservation. This process is called *data migration*. There are three levels of data migration:

- *Replication* — Straight copying of data with no alteration in physical format, logical format or information content. An example would be copying a tape to the same type of tape.
- *Repackaging* — Migration wherein the physical or logical packaging surrounding the content information may change but the information content is unaffected. An example would be copying a reel-to-reel tape to a newer type of cartridge tape.
- *Transmutation* — Migration where the information content changes in format but (ideally) where information content is preserved. An example would be converting character-coded data from ASCII to a newer format like Unicode (which includes ASCII as a subset).

The fundamental rule of data migration is that information content should be preserved if at all possible. For example, an image with 16-bit pixels probably cannot be changed to have only eight bits per pixel, because information would then be lost. On the other hand, the image could be compressed using a lossless compression algorithm with no diminishment in the information content. The archive's operating policies, along with the Submission Agreement, define what constitutes "preservation of information" for a particular AIC.

Typically the archive will retain all of the preservation information that was supplied in a SIP. This information will usually be attached to the pertinent AIUs and AICs and maintained by the Storage Entity, but some of this information may be maintained elsewhere in the archive to facilitate the management and retrieval of information. For example, if a company archives the CAD drawings for its products, it may keep a database that identifies each product by type, model name and number, dates of manufacture, and so on. This *catalog information* can make it easier for engineers to locate drawings of interest within the archive. It is up to the archivists to determine the kinds of catalog information to maintain; however, the community that an archive serves may have certain standards or guidelines regarding catalog format, content, or access methods that should be followed to allow users to locate information that may be distributed across several sites.

Up to this point our scenario has discussed the submission of information as SIPs and the

maintenance of information within the archive as AIUs and AICs. We turn our attention now to the retrieval of archived information. The chief reason that SIPs, AIUs, and AICs all include preservation information is to assure that the content information can be retrieved at a later (possibly much later) date. Typically the retrieval of information from an archive is a two-step process:

- First, the information of interest must be located and ordered. In the OAIS functional model, this is supported by the *Access Entity.*
- Second, the information must be extracted from storage, packaged, and delivered. These functions are performed by the *Dissemination Entity*.

As an example of this process, consider a physician who is doing research on a certain type of tumor. She has a list of patients who were diagnosed with this malady and she wants to see if the hospital archive has their CT scans, and if so she wishes to retrieve them. She connects to the hospital archive through the Access Entity which begins by validating that she is authorized to access this information. She then enters a query to see if there is information in the archive for any of the patients on her list. Different archives will handle queries like this in different ways. The Access Entity may have to retrieve from storage an AIC that holds all CT scans and search through all of the collection's AIUs looking for those that qualify. Alternatively the archive may have a catalog that provides a summary of each scan, including the patient's name. Having such a catalog would greatly expedite the search in our example, since scans of interest could be located without having to sift through the archived information.

Once the full set of qualifying CT scans has been identified, the Dissemination Entity takes over to transfer the results to the consumer. There are several options for disseminating the information to the consumer: the scans could be recorded onto a digital tape and sent through the mail; or they could be transferred over an electronic network.. Regardless of the method of distribution, a *Dissemination Information Package* is assembled. A DIP comprises the content information to be distributed — in this case the CT scans — along with some or all of the supplemental information that the archive contains on the scans. Just as archivists determine the mapping from SIPs to AICs, so too the archivists determine how to map from AICs to DIPs. First, the information in a DIP may be in the same format in which it is stored within the archive or it may undergo some reformatting —repackaging or transmutation, as discussed above — before it is sent to the customer. Second, a DIP may include the information from a single AIU or a collection of AIUs extracted from one or many AICs. In our example, only a small fraction of the total number of CT scans in the archive actually fits the consumer's query, so the archive prepares a DIP containing only the desired subset. Alternatively, if the archive determined that the desired scans resided on a single computer tape (along with many other non-qualifying scans), the archivists could simply choose to replicate the particular tape and send it the researcher along with package information that would specify how to locate the scans of interest on the tape.

ANNEX A. SCENARIOS OF EXISTING ARCHIVES

A.1 PLANETARY DATA SYSTEM ARCHIVE

I. DOMAIN

Domain and Consumers. The Planetary Data System is chartered to provide data archiving services, data access and expert help to the NASA-funded planetary science community. The PDS is a distributed system with a Central node at the Jet Propulsion Laboratory and discipline nodes (imaging, geosciences, atmospheres, planetary plasma interactions, small bodies, rings) located at universities around the country. The early focus has been on restoring historical mission data and has produced several hundred CD-ROM volumes containing about 80 per cent of the important planetary data archives. There has been an increased emphasis on providing access to the general public for educational outreach over the past several years.

Data Producers. Planetary data sets originate with NASA flight project data management and science teams (new data, some restorations), individual scientists (newly processed or value-added data) or via the PDS discipline nodes (restorations and value-added data). At least 50 per cent of the PDS resources have been devoted to restorations over the past seven years, with several more years of work needed to capture all historical data.

II. INGEST PROCESS AND INGEST INTERFACE

The PDS has developed a very formal interface with the major data producers (flight projects). This interface is documented in the Data Preparation Workbook and involves substantial interaction between node personnel, data engineers and project representatives. A Project Data Management Plan, signed by the PDS project manager provides the basic project data description and agreement to deliver to PDS. Since about 1993 all NASA announcements for Planetary investigations or analysis require that all data generated be delivered to PDS in conformance with PDS standards.

Submission Agreements

Projects provide a Project Data Management Plan. Sometimes a more specific document, the Archive and Transfer Plan, supplements the PDMP, providing extended product documentation and a schedule of deliveries.

Individual scientists can propose to be "data nodes" and receive funds from a PDS discipline node for preparing restored or value- added data sets for inclusion in the archive. There is no formal submission agreement for data nodes.

The PDS discipline nodes each maintain a list of outstanding restorations. These are worked-off based on their priority within discipline. At some point this list will be completed and only new project or data node data sets will be ingested into PDS. There is no formal

agreement associated with discipline data restorations.

Each data set that is identified for ingest in PDS is assigned to a Central node data engineer. It is the responsibility of the data engineer to see that all archiving steps are completed. The archiving steps are called out in the PDS Data Preparation Workbook.

Typical Data Delivery Session. Typically a delivery session will consist of a single data set contained on one or more volumes of CD-ROM or CD-Recordable media. A data set is defined within PDS to be a group of homogenous data granules at the same data level (raw, decalibrated, reduced) which differ only in time of acquisition and major category of target body. For example, the images of Jupiter taken by both Voyager spacecraft comprise a single data set. The standard process includes up-front negotiations between PDS and the provider; the production of test products which are evaluated in the peer review; revised final test products which are validated by the data engineering staff at the central node; approval and production of CD-ROM volumes; distribution by the appropriated discipline node or the central node; entry of the data set into the PDS central catalog; and entry of the data set into the NSSDC ordering system.

Transformation Process. In most cases the original data formats are maintained when data is brought into PDS. This allows existing software tools to continue to be used with the data. Much of the data preparation involves carefully documenting the data format and preparing metadata (granule labels, index files and catalog templates).

Validation. Validation is generally performed as part of the peer review of a product or by using validation tools. In some cases (for example, Magellan), the project develops its own internal validation process. The main validation tool of the PDS is the Volume Verifier. This program is run by the Central Node data engineers on each product delivered from a project or a data restoration. It validates the format and content of all product labels, and validates data files using checksums.

Security. The only area where any special security issues exist involves the receipt of proprietary data. Some projects have one-year proprietary periods before data in released to the science community. The PDS policy is to avoid receipt of any proprietary data sets during the proprietary period.

III. INTERNAL FORMS

The PDS has developed standards for documenting data sets (templates) and individual data products (PDS labels) using a keyword=value label system called the Object Description Language (ODL). Recommendations are also provided for volume organization and data product formatting to optimize the utility of resulting data products.

The PDS standards are specified in the PDS Standards Document. Standard documentation requirements include templates describing the data set, instrument, mission, etc. These templates are included on data volumes and also entered in the PDS high-level catalog. Standard terminology is maintained in the Planetary Science Data Dictionary, which is

jointly maintained by the PDS and the multi-mission ground data system. The metadata values for new data products are carefully compared with the PSDD and existing values used wherever possible. Additions are made to the PSDD to add new standard values to accommodate new data sets and when justified new keywords are added to the PSDD. Data products can have specialized metadata values which are not cataloged in the PSDD.

The PDS product labeling system is flexible enough to allow nearly any data structure to be described. Labels can be attached to the beginning of the data file or detached in a stand-alone text file which points to the data file. In some cases a single label file is used to describe multiple data files. Detached labels can be used to describe data stored in other formats (FITS or HDF, for example). In cases where complicated raw telemetry formats are stored the Software Interface Specification (SIS) for the product is included in lieu of descriptive labels.

Archive Volume Components

An archive quality data set is required to contain the following components.

AAREADME.TXT	- Text summary of data contents.
VOLDESC.SFD	- Standard volume label.
VOLINFO.TXT	- Text description of data contents.
CATALOG	- DATASET.CAT, MISSION.CAT, INST.CAT
INDEX	- ASCII index for each granule on the volume.
SOFTWARE	- Software needed to interpret/display the data.
CALIB	- Calibration data sets.
BROWSE	- Browse products for this volume.

Peer Review. All restoration and data node produced data sets are required to undergo a peer review before acceptance as archive products. Products produced by flight projects do not go through a formal peer review process. In general there is ongoing negotiation between the data engineer or the discipline node staff and the data producer. The peer review team consists of a number of scientists familiar with the data set, the discipline node leader and one or more data engineers. All product documentation and sample products and software are supplied to the peer review group for evaluation. The peer review group determines the adequacy of documentation and quality of the data products and either approves the product or provides a set of liens which must be fixed prior to approval. The PDS nodes and data engineers have access to a Volume Verifier tool which aids in validating the quality of an archive volume. The volume verifier checks internal checksums, verifies that the index contains entries for all data products and validates the volume templates as well as the descriptive keywords supplied for each product.

Delivery Media. Discipline restorations and data node products are recorded on CD- ROM or CD-recordable media as a standard practice. Flight projects are urged to provide archive quality products on CD media but may not be able to due to funding constraints. Products delivered to PDS on magnetic tape media are assigned to the PDS restoration queue. It is the

goal of PDS to convert all data sets to CD-ROM or CD-recordable media which is replicated at a separate geographic facility. This separate facility is generally the National Space Science Data Center (NSSDC) at Goddard Space Flight Center.

IV. ACCESS

Nearly all access to PDS data sets is via the CD-ROM volumes which are distributed to the entire research community. Large discipline node data collections including a substantial volume of CD-ROM data are accessible via the Internet. Several of the discipline nodes have developed on-line retrieval systems customized to meet the needs of their discipline scientists.

Finding Aids. The Pilot PDS devoted substantial resources to designing a central catalog system and distributed query and processing capabilities at the discipline nodes. These efforts were largely dropped as the Planetary Data System focused on data restoration rather than data access. In general, most of the user community already had home grown tools for data analysis and were most concerned with getting access to the data sets. The growth of the user community due to Internet and increased usage of CD-ROM readers has spurred to prototype a more consistent finding aid. The PDS Navigator has been developed for selecting images from the Clementine mission. It includes three components, a forms-based traditional database retrieval capability, an image-based retrieval and a text-based retrieval.

Security. The high-level PDS catalog can be accessed via a group account. Most of the data access services at the discipline nodes require the user to obtain a valid account on the node computer.

Customer Service/Support

The order function of the PDS is distributed. Data inventories are kept at NSSDC, the PDS central node and at each discipline node. In general each site serves a special group of users:

- discipline node - members of the NASA funded discipline
- central node - other NASA scientists and engineers, other agencies
- NSSDC - other scientists, agencies, public and foreign users.

The PDS Operator at the central node handles requests for PDS documentation or standard data products. The discipline nodes handle data requests from within their discipline and also provide expert help in the utilization and interpretation of the data. Access to tools is also provided.

V. DISSEMINATION

The vast majority of data dissemination is done via CD-ROM disc. Several hundred copies of over 500 titles have been distributed to date.

Subscriptions. Nearly all PDS distribution is done via subscriptions or standing distribution

lists. It is the responsibility of each discipline node to maintain a distribution list for its discipline scientists. This list determines the order amounts for most CD-ROM titles. The central node maintains a distribution list for engineering and management personnel and for other external recipients (reciprocal distribution, software developers).

Media/NetworkUse. Nearly all final products are delivered to the user community on CD-ROM. Archival products that need not be widely distributed are stored on CD-Recordable media, with a duplicate copy provided to the NSSDC. Most PDS data is available for downloading via anonymous ftp connection to a large CD-ROM jukeboxes at the central node and the imaging node.

Data Manipulation. Each discipline has a suite of government developed analysis tools which can be applied to the discipline data sets. These software packages are available for UNIX workstations or VAX VMS platforms. Several nodes provide the user a menu of processing functions that can be performed on selected data and will carry out requested processing and provide the results electronically or via media. The most widely used commercial tool is IDL.

Pricing Policy. The PDS distributes data to legitimate NASA researchers for no charge. There are no charges for on-line computer usage or data processing to NASA researchers. The NSSDC distributes CD-ROMs for $10 per volume.

Security. All PDS data sets are certified GTDA by the Department of Commerce and are distributable worldwide.

VI. SPECIAL CHARACTERISTICS

PDS has invested a substantial engineering effort in its common data dictionary, data standards and procedures for preparing archival quality data sets. By having these standards in place the PDS is able to demand better quality data sets of its data providers.

A.2 NATIONAL ARCHIVES AND RECORDS ADMINISTRATION'S CENTER FOR ELECTRONIC RECORDS

I. DOMAIN

Domain and Consumers

The Center for Electronic Records is the organization within the U. S. National Archives and Records Administration (NARA) that appraises, accessions, preserves, and provides access to federal records in a format designed for computer processing. NARA serves as the archives for the records of the United States federal government. Consumers for this data are as diverse as the electronic records they seek to access and range from individuals seeking to assert their rights to other government agencies to academic researchers, private consultants, media personnel, and a wide variety of other users.

Data Producers

Originally this data is produced (created or received) by agencies of the U.S. federal government (producers). The data may concern virtually any area or subject in which the government is involved. They may come from a variety of computer application such as data processing, word processing, computer modeling, or geographic information systems. They can include records made directly by government employees or indirectly through government grants and contracts.

Special Features

The most noted special feature of NARA's Center for Electronic Records is the diversity of the collection of more than 29,000 data sets from more than 100 bureaus, departments, and other components of executive branch agencies and their contractors and from the Congress, the Courts, the Executive Office of the President, and numerous Presidential commissions. A small portion of the data originally were created as early as World War II. An even smaller portion contains information from the nineteenth century that has been converted to an electronic format. Most of the data, however, has been created since the 1960s. The major types of holdings and subject areas include agricultural data, attitudinal data, demographic data, economic and financial statistics, education data, environmental data, health and social services data, international data, and military data.

Scientific and technological data already transferred to the Center include the National Register of Scientific and Technical Personnel; the National Engineers Register; the 1971 Survey of Scientists and Engineers; major portions of the National Ocean Survey's Nautical Chart Data Base; numerous Environmental Protection Agency series relating to pesticide use, hazardous wastes, and pollution abatement; the Nuclear Regulatory Commission's Radiation Exposure Information Reporting System; biometric data sets and epiemiological studies (such as the National Collaborative Perinatal Project) from the National Institutes of Health, the Centers for Disease Control, and the National Center for Health Statistics; and text from presidential commissions on Three Mile Island, coal, and the Space Shuttle Challenger Accident. While the Center's scientific and medical holdings are rich and varied they do not fully reflect the extent and diversity of federal activity in this area.

II. INGEST

The ingest process begins with producers (records managers and records creators in federal agencies) inventorying all electronic records and determining how long to retain the records for current agency business. The next step in the process is for the producer and NARA to develop a *Request for Records Disposition Authority*, Standard Form 115 (SF 115). Here information on the content, retention and disposition, and the availability and extent of documentation and related reports is listed in the context of the producer's business needs for the information. Data with continuing value are listed as permanent and the timing and frequency of their transfer to NARA is established. The producer submits the SF 115 to NARA for its review and appraisal. The Center for Electronic Records appraises electronic

records items on all SF 115s. Identifying permanently valuable electronic records for retention by NARA's Center for Electronic Records involves cooperation between NARA and the producers. Through the process of scheduling and appraisal, the Center identifies and selects the electronic records it judges to have enduring value. The Center evaluates electronic records in terms of their evidential, legal, and informational value and their long-term research potential. Some of the factors in this appraisal evaluation include estimation of past, present, and probable future research value within the context of the data's origin and current use and its impact on federal programs and policy. Administrative and legal value, as well as the potential for linkage with other data, may bear on the decision. Unaggregated microlevel data sometimes has the greatest potential for future secondary analysis. Once the Center determines the records have enduring value, it then determines whether the records should be preserved in electronic format.

Submission Agreements

The actual Submission Information Package (SIP) between NARA and the agency that creates or receives the data is a *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States*, Standard Form 258 (SF 258) accompanied by the data object(s) and sufficient documentation and descriptive information to use the data. The SIP transfers physical and legal custody of the electronic records from the producer to NARA. This agreement is the end product of the ingest process described above. The SF 258 also contains any restrictions on access to the data which conform with exemptions listed in the U.S. Freedom of Information Act. The Center enforces all legitimate restrictions on access. The Center also works with the producer to determine if any "disclosure-free" version of the data can be produced for consumers.

Typical Delivery Session

This inventorying, scheduling, and appraisal process specifies the data object(s) and related metadata and documentation to be transferred and establishes the timing and frequency of submissions. Specific instructions for how the data are to be organized and when they should be submitted are established in the *Code of Federal Regulations* (36 CFR 1228.188). All data should be transferred on either open reel magnetic tape, tape cartridges, or CD-ROM. The CFR sets the specific technical requirements in terms of format, block size, and extraneous characters. While the current regulations also require that all SIPs should be transferred in a software-independent format, NARA staff recognize that the research potential and utility of some data would be significantly reduced if they were transferred in such a format. In such cases NARA works with the producers to determine the best mode of transfer.

What are the Information Objects that are Delivered? Producers typically will transfer a series consisting of one or more data sets with the related documentation which minimally should include the record layout and codes, methodology statements, technical information about the data including number of records and size. Ideally, the SIP also includes associated analyses and reports. Increasingly agency-created metadata also is included. The majority of electronic records come as flat files of data; increasingly, however, text files and output from

data base management systems, and geographic information systems also are transferred.

What are Collections? NARA organizes all Archival Information Collections (AIC) on the basis of Provenance and Original Order. Provenance maintains the identity of an Archival Information Package (AIP) or an AIC and preserves as much information as possible about its origins and custodial history. Within NARA this is accomplished through the use of Record Groups which reflect the structure of the federal government and subgroups and sub-subgroups which place the AIPs and AICs within the producer's place within its agency. Original order argues maintaining the contents of an AIP or AIC in the order developed and used by the producer. This helps reveal the producer's organization and how it used the data objects and can provide additional information to consumers. For electronic records, "original order" is expressed in the logical structure of files and databases and in the indexing which the producer used. Within NARA the basic unit for arrangement and description is the AIC which can include a number of related AIPs.

What Descriptive Information is Provided? The extent and quality of the descriptive information provided by the producer varies from quite sketchy to extremely detailed. NARA staff attempt to flesh out the producer-created descriptors with AIC level descriptions, title list entries, abstracts, and Dessimination Information Packages (DIP) and to provide the descriptive information in a variety of formats to reach different consumers.

What sorts of Validation Objects are Provided? Producers are required to transfer metadata and descriptors adequate to access, process, and interpret electronic records. For formatted data files the DIP must include a record layout with appropriate field definitions and codes. It frequently also includes methodology statements, input documents, data entry instructions, processing directions, sample outputs, reports and analyses of the information and system manuals.

What Transformation Processes are Performed Prior to Storage

What Metadata is Created? The most extensive metadata product created by NARA is the DIP. In the Introduction, Center staff discuss the origin, creation, and administrative uses of the data object(s), list related objects that are or will be available, and discuss characteristics of the data that could cause problems for consumers based on initial validation processes. The DIP also includes sample printouts of the data and tables and reports related to computer validation of the data. NARA also captures metadata on record layouts, domains, ranges, and links between files in a metadata database as a byproduct of the automated validation process. Other metadata created by Center staff include AIC descriptions, formatted abstracts, title line entries, and collective descriptions which place the data in a broader context. The Center anticipates that increasingly metadata created by the producer will be part of the SIP transferred to NARA.

What Validation is Performed? The Center's initial accessioning procedures include creating a new master and backup copy of each data object on new certified media to ensure the best physical media for long-term storage. At this time Center staff perform automated comparisons of the data contents with the record layout and codes, and of the physical

structure including the number of records, blocks, and bytes. Staff also perfect the DIP to facilitate secondary use of the data.

Security

All data are maintained off-line with consumer access only to copies of the data. The master and backup copies are maintained in separate secure stacks at two different physical locations. Data which require additional security measures, for example Census data subject to restrictions imposed under Title 13 of the *United States Code* and national security classified information restricted under Executive Order, are afforded the appropriate level of protection. The Center is moving to provide enhanced access to selected data onsite by providing reference copies on a wider variety of media and by providing a broader range of services and output products. This may include use of vendors who can provide enhanced access to the holdings utilizing "value-added" services.

III. INTERNAL FORMS

How do you Store your Data? All master and backup copies are stored on newly certified class 3480 magnetic tape cartridges. Some of the holdings have not yet been migrated from nine-track, 6250 bpi open-reel magnetic tape. Data are received and stored temporarily on other media including diskettes, 4mm, 8mm, CD-ROM, and various removable hard drives, although not all of these media conform with regulatory requirements.

Migration (Data). Based on recommendations from the media manufacturers, the National Technology Alliance, the National Institute of Standards and Technology, and various standards organizations, the Center has been migrating its data to new class 3480 magnetic tape cartridge when each media unit is ten-years old.

Migration (Metadata). Metadata has been stored in a variety of formats depending on the original format transferred with the data. Traditionally most metadata existed in textual format. The metadata captured in the validation process is maintained in a relational database. There are no current plans for migrating from this format, although the metadata can be exported in flat file format. The Center has been encouraging data producers to create and transfer metadata in electronic form. Within the next fiscal year the Center hopes to begin scanning and digitally converting metadata so it can be preserved and provided in an electronic format along with the data.

Migration (Format). The *Code of Federal Regulations* requires data producers to transfer all data in ASCII or EBCDIC with all extraneous characters removed from the data except record length indicators or tape marks and blocked at no higher than 32,760 bytes per block for open-reel and 37,871 bytes for class 3480 magnetic tape cartridge. When CD-ROM is used they must conform to ISO 9660 standard and the data must be in discrete files containing only the permanent data. Additional software files and temporary files may be included on the CD-ROM. The CFR also requires all electronic records to be transferred in a software-independent format. The Center works with data producers who cannot meet those requirements to determine the most appropriate transfer and storage formats.

IV. ACCESS

What Finding Aids are Provided?

Information about the holdings are available in multiple levels of detail and by multiple sources as a way to provide various consumers with information about the Center's holdings. The least specific detail is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States* where electronic records series are described in the context of the larger holdings from a producer. Other collective descriptions include *Information About Electronic Records in the National Archives for Perspective Researchers,* General Information Leaflet 37, which also is available on the Center's homepage (http://www.nara.gov/nara/electronic), and a title list of data sets available on the Center's homepage and as a printout. Specific electronic records series descriptions were created as formatted metadata for a portion of the Center's holdings for inclusion in a proposed automated description data base which has not been implemented. The most detailed description for any data set is the DIP. Each DIP may contain a narrative describing the data file(s), the record layout and codes for the data, a methodology, sample input forms and questionnaires, annotations regarding the data validity, and a bibliography. The Center also has established an email site (cer@nara.gov) for queries regarding the Center's holdings and services.

Security.

All of the Center's holdings are maintained in environmentally controlled closed stacks which are accessible only by Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. The Center's national security classified data sets are in separate environmentally controlled stacks approved for the storage of classified information. All processing is performed in limited access processing rooms at NARA or at the National Institutes of Health computer center. Computer processing is done on closed systems which require both a registered logon and personal identification number or password to access the system. Researchers do not have direct access to any accessioned data. Presently they access copies of the data that they have purchased for their own use.

Customer Service/Support.

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The staff responds to both general and specific inquiries by telephone, letter, email, or in-person visit and fills orders for copies of specific data and their DIPs. The staff also provides information from records to respond to researcher requests such as for casualty records from the Korean and Vietnam conflicts. The staff also functions as a filter between researchers and the data producers when problems develop in understanding or interpreting the data. The staff develop a variety of informational material about the Center's holdings and services, much of which is available online.

V. DISSEMINATION

Do You Support Subscriptions?

The Center will accept a standing order (subscription) for electronic records that it receives on a regular, periodic basis from producers of the Federal government. Under current NARA regulations all subscriptions must be prepaid prior to shipment of the data.

What Media/formats do you use?

Currently the Center provides copies of data files on either nine-track open-reel magnetic tape or class 3480 magnetic tape cartridges encoded in ASCII or EBCDIC, labeled or unlabeled and written to the maximum block size requested. The Center also can provide an exact copy of records in nonstandard formats, if the agency transferred them this way, but it cannot validate or verify the contents of these files. In the past these other formats have included packed decimal, zone-decimal, binary, National Information Processing System (NIPS), Statistical Analysis Software (SAS), Statistical Package for the Social Sciences (SPSS), or OSIRSIS. The Center recently expanded its media options to include diskettes for smaller data sets and CD-ROM. On-line transfer of data remains a more distant goal.

What Transformation (Value Added) is Provided?

The Center currently preserves data as received from the producers; it does not routinely provide extracts from the data or other value-added services beyond computer validation of the data contents and enhanced documentation. Planned enhancements provide for value-added services including extracts from the data.

Pricing Policies.

The Center uses a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently, the charge for an exact copy of all data on a input cartridge or reel, regardless of the number of data sets on the media is $80.75 when copied to a class 3480 magnetic tape cartridge and $90.00 when copied to a nine-track open reel magnetic tape or a CD-ROM. The Center charges an additional $24.50 for each additional data set or file added to a media and $7.50 for each subsequent magnetic tape cartridge and $17.00 for each subsequent open-reel magnetic tape. Paper reproductions cost $0.25 per page.

Security.

The same security considerations developed in relation to Access apply to Dissemination. The Center's national security classified data is made available only to researchers who have both the appropriate security clearances and the appropriate need-to-know. Other restricted data are made available only with prior written approval of the creating agency or under the terms of the restrictions which must be supported as a legitimate exemption under the Freedom of Information Act.

VI. SPECIAL CHARACTERISTICS

NARA's Center for Electronic Records has a diverse collection which reflects the diverse activities of the federal government. The staff shape the holdings through the process of scheduling, appraisal and accessioning. Currently, the Center acquires less than one percent of all federal records created in an electronic format. The timing of the transfer of electronic records from the creating federal agency to NARA is negotiated with the creator to ensure that the records are available for agency use for as long as necessary for current business and that they are transferred to NARA as soon as practicable to ensure their long term preservation for secondary use. NARA is the only federal agency with an explicit archival mandate for Federal records and thus the only Federal agency that preserves and provides access to a wide range of historically valuable records for the indefinite future. As such it is an archives of last resort for the electronic records of some federal agencies which undertake an active data dissemination function while there is a researcher interest in the data but whose mandate ceases or may cease once the demand wanes or ceases.

A.3 LIFE SCIENCES DATA ARCHIVE

I. DOMAIN

What is the domain and who are the customers of the Archive and who are the producers of the data? What are the special features of this archive?

The Life Sciences Data Archive (LSDA) is responsible for collecting and disseminating data of NASA funded Life Sciences space flight investigations. There are two general goals for NASA space life science research; one, to find counter measures to problems encountered by human bodies as a result of space flight, and two, to broaden the understanding of the effect of gravity on living systems. The LSDA's primary customer is the life sciences research community, but it is also used by students, educators and the general public. The data archived in the LSDA is produced by both intramural and extramural investigators funded to perform flight experiments through NASA grants. It is anticipated that the archive may grow to include data from investigations which are completely ground based.

The LSDA is a distributed archive with responsibilities distributed to LSDA Nodes at various NASA Centers and Projects with life sciences activities. There are the LSDA Project Nodes (ARC, KSC, & JSC) which are responsible for the actual collection and cataloging of data, and there is a LSDA Data Distribution Node (Central Node) which is responsible for dissemination of the data to the public.

The LSDA contains animal, plant, and human space flight data. This archive is notable in that it contains a unique collection of data describing, in considerable detail, biology experiments carried out in space by NASA over the past thirty years. The nature of the data is highly varied and spans many life science disciplines.

The LSDA is also unique in that it provides both digital and non-digital information. The non-digital data may be either reproducible or non-reproducible. Examples of reproducible,

non-digital data are video and audio tape. An example of non-reproducible, non-digital data is a biomedical sample.

II. INGEST PROCESS

Submission Agreements

There are two major types of data producers; the NASA Flight Project offices that design hardware and manage the experiment, and the NASA funded Principal Investigator.

To acquire data from the NASA Flight Project offices, the LSDA Project Nodes work closely with them to acquire data during flight operations. The LSDA assists the NASA Flight Project Offices in distributing this data to the Principal Investigators and gathering it as an archival product. As the LSDA is relatively new (1993) there is also retrospective archiving of past missions being done on a funding available basis.

To acquire data fro the NASA funded Principal Investigator, there are a couple of methods of data collection currently being used depending on the "age" of the experiment. For previously flown experiments (prior to 1994) there is an informal submission agreement between the LSDA and the PI's that is based on cooperation, and is not binding. For experiments being selected for flight (after 1994) the funding agreements include a contractual stipulation that the PRINCIPAL INVESTIGATOR must supply the LSDA with raw data, analyzed data and a final science report.

These funding agreements are finalized when proposed investigations are selected for flight. At this time the PIs are sent a letter informing them, that upon acceptance of funding they will be responsible for delivering the data collected as part of their investigation in a form usable by the sciences community one year post flight.

After a one year proprietary period, submission of data to the LSDA begins. To assist in its submission, the LSDA Project nodes send the PRINCIPAL INVESTIGATOR a Data Inventory package. The PRINCIPAL INVESTIGATOR fills out the data inventory forms and returns them to the LSDA Project Node. The Project Node then contacts the PI to begin data submission. In order to clarify the "usable form" requirement throughout the entire LSDA project, the LSDA is in the process of developing a post flight data reporting handbook which explains exactly how the data should be provided to the archive.

Typical Delivery Session

A typical submission information package (SIP) consists of two parts; 1) the Data Inventory forms, and 2) actual data. The inventory forms are considered CI and contain information about the data types (i.e. physical samples, hardcopy/photographic/video, computer files, other formats) and Data Sets (i.e. title, description, treatments, parameters measured, research subjects and Ids, date/period of collection, collection location, analysis phase, comments). The actual data consists of physical biospecimens, spreadsheets, final science reports, published articles, procedural documents, photographs, video tapes, analog tapes, digital or

printed images, and other types of digital data files (i.e. HRM).

Upon receipt by the LSDA the CI will be cataloged and Archival Supporting Information added, including; experiment and mission ID, Principal Investigator and Co-Investigators name, and other linking information.

Collections

The Archival Information Unit's (AIU) are compiled per investigation, i.e., all AIUs for a single experiment make up an Archive Information Collection (AIC). During a typical ingest session each SIP is cataloged and Archive Supporting Information is added. This CI is entered into a database comprised of LSDA approved fields and uses valid values whenever possible. The Archive Supporting Information is developed by the LSDA personnel at the LSDA Project Node responsible for obtaining the data. The Archive Supporting Information provides layers of metadata for the data collection that describe the experiment, mission, hardware, personnel, sessions, biospecimen, and research subjects from which the data was collected.

It is anticipated that future uses of the archive will involve the creation of AICs based on discipline or measured parameters.

Transformation Processes

In most cases a set of data is kept in its original submitted form. Exceptions to this case include data submitted on outdated media requiring transfer to current media. As little transformation as possible is performed on the data at ingest in order to keep costs down and to insure the integrity of the data. There are some instances where the data has been collected in an application format that is not widely available and in this case the Project Node will transform the data into a more commonly accessible format. (e.g. spreadsheets created in Supernova are translated to MS Excel).

After the LSDA Project Node enters/creates this CI or metadata for the data collection and the individual data elements, the information goes through a validation process. This post-entry validation is accomplished by a second check of the data by the LSDA Project Node Manager. Content validation is further ensured by sending the completed catalog entries to the data originator (Principal Investigator, Flight Project Offices) for verification. The Principal Investigator reviews the information, makes corrections or additions and sends the information back to the Project Node. Edits are then made to the records and the information is once again printed and sent to the Principal Investigator. This process is repeated until the Principal Investigator is satisfied that his experiment data is accurately represented. At this point the Principal Investigator signs and returns a verification letter to the Project Node. The catalog or metadata is now ready for review by the LSDA Project Scientist before it is placed in the public record (via Web Site). The LSDA Project Scientist will review the data for overall form and cogency. Do you want to put anything in here about the LSDA review cycle or is review by the LSDA Project Scientist a sort of catch all term?

Security

The LSDA has strict security measures for data from human subjects which require sensitivity and secure handling due to the Human Data Privacy Act. Human data when received is coded to protect the identity of the crew members. Security procedures include keeping the data on magneto-optical disks stored in a locked file cabinet in a cipher locked room.

Overall security procedures stipulate that all digital data are backed up on a daily basis with off-site storage. Access to on-line servers is controlled through the use of password and/or address port filtering. Only data that is fully validated and approved for release is placed on publicly accessible servers.

III. INTERNAL FORMS

Storage

The LSDA back-up and storage procedures vary between LSDA Node types. Currently, the LSDA Data Distribution Node resides at Johnson Space Center. The LSDA Master Catalog and on-line data reside on a Microsoft SQL Server. These are backed up to tape daily. At the LSDA Project Nodes most of LSDA's data and metadata are stored on magnetic disks and backed up to tape. Long term storage is provided on CD-ROM.
Biospecimens are stored in -80 degree freezers.

AIUs are stored as a piece of CI (a spreadsheet, word processing document, strip chart or biospecimen) with the archival preservation information stored in a database record. The CI can only be, easily, accessed through the database record with it's descriptive information and the CI storage directory information. These AIUs are linked, through the database, into AICs via an Experiment Number. A space life sciences experiment is, in this sense, an AIC. It is a collection of tens or hundreds of AIUs.

Migration

The LSDA migration process is still in a developmental phase but there is some ongoing data migration. LSDA Project Nodes are in the process of converting information on outdated media (RA60's, RL02's) to CD-ROM format.

Migration of application formats (e.g. MS-Excel) and in particular, version changes, is an area of concern. The cost of continually updating all LSDA spreadsheets to the current version is prohibitive and storing and making available the application is also expensive and complicated. A universal read only format such as Adobe Acrobat might be the solution, but ~~it~~ is a proprietary format and it's life span is an unknown.

IV. ACCESS

Finding Aids

Access to LSDA information and data is handled through the LSDA Data Distribution Node at JSC. Users enter the LSDA through the World Wide Web (WWW) and search/retrieve information via the Master Catalog. The Master Catalog is a relational database with a WWW forms interface and allows users to search Archive Supporting Information across experiments and Missions to find data that meets their search criteria. Users can search within ten information groups; Experiments, Missions, Data Sets, Hardware, Documents, Personnel, Specimen or Subjects, Data Collection Sessions, Biospecimens, and Images.

There is currently no method available for searching data at a 'sub-AIU' level. The AIU record contains a considerable amount of detailed, searchable, data so that a collection of AIUs could be found for a particular manual sub-search.

*Security***.** The LSDA does not have any special security concerns for access to the public Master Catalog information and non-human digital data. It is freely available to anyone on the WWW. However, the human flight experiment data is subject to the Human Data Privacy Act, and therefore, security measures are required to control access to this data. The policies and procedures for access to the data are currently being developed.

*Customer Service/Support***.** The LSDA provides user support for questions and problems concerning the Master Catalog (on-line data request system) and for questions about the data being provided. The primary means of user feedback and support is through the LSDA Data Distribution Node. Questions are addressed to the LSDA through on-line “What do you think?” links located throughout the system. From these links a WWW forms interface allows users to submit questions. Specific questions about the data are currently addressed by the NASA Life Sciences Acquisition Scientist and the LSDA Program Scientist. Questions which can not be answered by these individuals are forwarded to the LSDA Project Node which collects the data. In some instances questions are forwarded to the Principal Investigator or NASA Flight Project Office who provided the data.

V. DISSEMINATION

Subscriptions. The LSDA does not support subscriptions since the publicly available Master Catalog is accessible to all users. The LSDA data is located using a catalog on the WWW. Most data is disseminated to the user through links in the catalog to an anonymous FTP site from which the data is downloaded. This means of data dissemination is, therefore, tightly linked to the data “finding” process. If data are in non-digital format, but are reproducible (i.e., hardcopy documents, or log books), users may request them through on-line ordering forms available in the Master Catalog. The requested information is reproduced via photocopying and shipped US Mail to the requester.

There are discussions about an update notification service to be offered in the future.

Media/Formats. The LSDA contains unique non-reproducible pieces of data such as microscope slides and space flight biospecimens. These unique resources are provided to a requester after a scientific proposal has successfully undergone peer review. Biospecimens,

once disseminated, are used to produce original data which is then ingested into the archive.

*Transformations***.** Currently, the LSDA does not provide many value added services. The data is stored and disseminated as provided by the data producer. Data are available in raw and summarized form. These summarized data are provided by the data producer. LSDA does convert data which are received in a non-standard format to a more usable form. Currently there are no data analysis tools available through the LSDA. However, the LSDA Project Nodes do ensure that all data sets have the minimum amount of information needed for understanding (e.g. explanation of all column headings are provided for the spreadsheets, etc.). These are the only "value added" processes that come from the raw data.

Security. Since the Master Catalog and non-human data in the LSDA are available to anyone on the WWW there is no special security in place for the dissemination of this data. However; the human flight experiment data is subject to the Human Data Privacy Act, and therefore is not openly available to users. Limited dissemination of human data will be allowed using policies and procedures that are currently being developed.

Pricing Policies. LSDA data that has been verified and cleared for release is available to the public, free of cost, through the Internet. If significant requests are generated for hardcopy documents, a processing fee for copying the document may be charged. As yet this has not been determined. In the future CD-ROMs with data may be generated. These CDs will be priced in order to recoup production and distribution costs.

A.4 NATIONAL COLLABORATIVE PERINATEL PROJECT (NCPP) 1959-1974

I. DOMAIN

Domain and Customers

The National Collaborative Perinatal Project was a multi-institutional, multi-year study of pregnant women and the children born from those pregnancies to provide baseline information useful for later determining the causes of neurological diseases which appeared in a portion of the studied population. The data came from medical histories, examinations, and observations. The records also contain socioeconomic, family history, and family health information. The data are used by a variety of medical and other researchers.

Data Producers

The predecessor to the U.S. National Institutes of Health's National Institute of Neurological Disorders and Strokes (NINDS) began the National Collaborative Perinatal Project in 1958. Fourteen university-affiliated medical centers across the United States participated in the study. Between 1959 and 1965 each cooperating medical center collected information on between 300 and 2000 pregnancies each year for a total of 55,908 pregnant women utilizing their clinic services. This represented between 14% and 100% of the women utilizing these services depending on the sampling rate employed at each clinic. The final population was reduced to 39,215 due to miscarriages prior to twenty weeks, 445 multiple births, exclusion

of subsequent or repeat pregnancies, and deletion of incomplete records due to women withdrawing from the study prior to its completion. The children were given neonatal examinations and follow-up examinations were through eight years of age. The last examinations were conducted in 1974. The computer data files resulting from the research that NINDS transferred to NARA consist of approximately 6,200,000 records organized into a Master File, a variable file, and eighteen work files, one of which consists of thirteen distinct data files.

Special Features

The Collaborative Perinatal Project was a longitudinal multi-disciplinary research effort which sought to relate the events, conditions, and abnormalities of pregnancy, labor, and delivery to the neurological and mental status of the children of these pregnancies and their siblings through eight years of age. The study sought to link any later appearance of cerebral palsy, mental retardation, learning disorders, congenital malfunctions, minimal brain dysfunction, convulsive disorders, visual abnormality, or communicative disorders to patterns during the perinatal period in order to develop strategies for prevention and intervention. The sample population is large enough so that statistically significant numbers of such disorders would appear in the children. Study of the records relating those children could result in the development of predictive factors and possible preventive care or intervention actions which could reduce future incidence rates.

The data are available in two formats: microfilm of the individual case files for the mother and child of approximately 270 pages per case file and the computer data files. Access to the microfilm and two of the computer data files is restricted because they contain personal identifiers. The National Archives has created a public use file for the Master File and Work File 16: Serum Specimen Inventory.

II. INGEST

The ingest process for transferring any federal agency records to NARA begins with the agency identifying the records and assessing their potential evidential, legal or research value. The next step is for the agency to develop a Standard Form 115, *Request for Records Disposition Authority*, and submit it to NARA. NARA staff then appraise the records in terms of their evidential, legal, and informational value and their long-term research potential. NARA and the creator then establish a transfer date, negotiate any restrictions on access, and initiate the ingest process.

Ingest for the NCPP computer data was a two phase process. In Phase One, from 1958 through 1974, NIH's NINDS funded the project and the cooperating institutions conducted the research. Contractors accumulated the original examination records, created the consolidated case files, microfilmed the records, normalized the data, and developed the Master File, an extract file of frequently used variables, and special files such as "refined diagnoses". The data were stored on 23 reels of magnetic computer tape recorded at 1600 bpi. Prior to 1980 the data were available only to NINDS, the cooperating hospitals, and selected government researchers.

In Phase Two NINDS developed the documentation necessary for more generalized use of the data and negotiated a submission agreement, including access provisions, with NARA. Since the nonreleasable data could be made anonymous through creation of a Public Use File, the producer and NARA worked on transferring the data files first.

Submission Agreements

NARA and NIH executed the U.S. Government's standard transfer form, Standard Form 258, *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States,* in mid-1985. This transferred legal custody and preservation responsibility to NARA. A similar agreement for the microfilm examination records was executed in 1990 after NARA and NIH resolved the privacy and access concerns and NARA developed a statistical research form.

Delivery Session

The delivery session was a single transaction in which NIH provided NARA with copies of the 23 reels of magnetic tape containing the NCPP and the related documentation consisting of seven volumes containing the background of the study, the sample, data collection and data processing overviews, record layouts and coding for each variable, sample forms, and a bibliography of all published research through 1985. NINDS transferred the 8000 rolls of microfilm containing the examination records in 1990.

Transformation Process

NARA has maintained and preserved the original data format. The data are in a hardware and software independent EBCDIC format which facilitates wide researcher access. All data were copied to new nine-track open-reel magnetic tape when received in 1985 and are migrated to new media every ten years to ensure long-term preservation. The more than 7000 pages of documentation are available in both a page format and in a microfiche format on 75 fiche. The documentation has not been scanned or digitized.

Validation

Validation was performed as part of quality control throughout the life of the NCPP. Extensive use of the data during the life of the project (1958-1974) and its use by NIH approved researchers (1958-1985) provided a second de facto validation. NARA also validated sample portions of the data at the time of ingest. Continuing researcher use also validates the data contents.

Security

The computer data is maintained in environmentally-controlled closed stacks which are accessible only to Center staff. Master and backup copies of the data are stored in separate

vaults in separate locations to facilitate disaster recovery. NARA has created Public Use Files of the restricted data files to prevent unauthorized access to personal and medical data.

What Descriptive Information is Provided?

NARA has prepared multiple levels of descriptive information for the NCPP. These range from entries for each data file in the Center for Electronic Records' Title List, an abstract entry for the series, a series description, a full documentation package, to a series-level entry in the three-volume *Guide to Federal Records in the National Archives.*

What Validation Objects are Provided?

During its active life the NCPP established and used elaborate data collection, input and verification procedures. Extensive use also validates the information. NARA's routine transfer and storage procedures also validated the data. The extensive seven-volume documentation includes record layouts and codes, methodology statements, data analyses, and a bibliography of research use.

What Transformation Processes are Performed Prior to Storage?

What Metadata is Created? NARA staff supplemented the documentation with an abstract and introduction discussing the origin, creation, and uses of the data, including an explanation of restrictions on access and the characteristics of the Public Use File.
What Validation is Performed? NARA's accessioning and storage procedures included creating a new master and backup copy on new certified magnetic media and creating a Public Use File of the two restricted data files. Sample portions of each data set also were verified against the documentation.

III. INTERNAL FORMS

Storage. NARA maintains separate sets of the master and backup copies of the data and the Public Use Files on newly certified 3480 class magnetic tape cartridges.

Migration (Data). NARA migrated the NCPP from 23 nine-track, 1600 bpi open-reel magnetic tapes it received in 1985 and stored the data on seven nine-track, 6250 bpi open-reel magnetic tape. NARA migrated the data to four 3480 class magnetic cartridge when the media was ten years old.

Migration (Metadata). NCPP metadata is available in textual (7000+ pages) and microfiche (75 fiche) forms. There are no plans to scan or digitize the text.

Migration (Format). The data currently are encoded in EBCDIC with all extraneous characters removed. There are no plans to migrate the format at this time.

IV. ACCESS

What Finding Aids are Provided?

Information (of varying detail) about the NCPP is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States*, where the records are described in the context of the larger holdings of the National Institutes of Health; in *Information About the Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37; in *Title List: A Preliminary ad Partial Listing of Data Files in the National Archives and Records Administration*; and in the documentation package for NCPP. Much of this information is available on the Center's homepage (http://www.nara.gov/nara/electronic) or by posting an enquiry to the Center's e-mail site (cer@nara.gov).

Security

NCPP, like all of NARA's holdings, is maintained in environmentally-controlled closed stacks which are accessible only by authorized Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. Researchers do not have direct access to the data. Presently they acquire copies of the data on a cost-recovery basis permitting indefinite use of the data for their own purposes.

Customer Service/Support

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The reference staff responds to inquiries by telephone, mail correspondence, e-mail, or in-person visits. They fill orders for copies of all or part NCPP and the relevant documentation. The staff also function as a filter between researchers and the NINDS when problems develop in understanding or interpreting the data.

V. DISSEMINATION

Do You Support Subscriptions?

NARA's Trust Fund is willing to establish accounts that allow researchers to acquire data that is transferred on a recurring basis. Since the NCPP stopped collection data in 1974 there is no need for a subscription for this data.

What Media/Format do you use?

Copies of the 32 data sets comprising the NCPP are available on seven nine-track open-reel magnetic tapes, six 3480 class magnetic tape cartridges, or two CD-ROM.

What Transformation (Value Added) is Provided?

The NCPP is provided as received from NINDS. NARA has created Public Use Files for the two data files containing personal identifiers in conformance with the Freedom of

Information Act and NARA restrictions on access to records whose release might result in unwarranted invasion of personal privacy.

Pricing Policies.

Electronic data sets are available on a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently the charge for an exact copy of all NCPP data on a storage reel or 3480 class cartridge is $80.75 when copied to a 3480 class magnetic tape cartridge and $90.00 when copied to a nine-track open-reel magnetic tape. Copies on CD-ROM are $90.00 for the first file and $24.50 for each additional file written to the CD-ROM.. Paper reproductions cost $10.00 for the first 20 pages and $5.00 for each additional block of 20 pages. Microfiche reproductions cost $2.10 per fiche.

VI. SPECIAL CHARACTERISTICS

The National Collaborative Perinatal Project was a prospective study. NINDS expended more than $200 million over two decades to collect information on more than 58,000 pregnant women and their children at fourteen cooperating institutions. It is unlikely that a study of this duration and magnitude will be repeated. The data continue to constitute an important resource for biomedical and behavioral research in many areas of obstetrics, perinatalogy, pediatrics, developmental psychology and other fields.

A.5 ARCHIVE SCENARIO FOR THE *CENTRE DES DONNEES DE LA PHYSIQUE DES PLASMAS* (CDPP)

I. DOMAIN AND CUSTOMERS

The CDPP (*Centre des Données de la Physique des Plasmas* - Center for Data on Plasma Physics) is a new service currently being set up. It has been developed to ensure the Long-term conservation and availability of natural Plasma Physics data (magnetospheric plasma, planetary plasma etc.) for the international scientific community. More specifically, the data concerned is from either ground-based or space-flown experiments in which France has participated or wholly directed. The CDPP is designed around two principal components:

- A Technical Activity segment, located on the premises of the French space agency, CNES, mainly in charge of developing and maintaining the archive system. The latter has the following functions: addition of data and metadata to the system, preservation of data and metadata, organization of search and product ordering facilities, and dissemination.
- A Scientific Activity segment, located at the CESR (*Centre d'Etudes Spatiales des Rayonnements* - Center for the Study of Space Radiation), a science laboratory near CNES. The CESR is in charge of all aspects relating to scientific knowledge of the data: validating data with its producers, ensuring that the data is useable by the scientific community, setting up added-value services etc. This Center is also responsible for developing a WWW server to present CDPP services, supplying educational information on Plasma Physics to the general public, and guiding users to access and dissemination functions.

The two complementary segments work closely together.

A number of associated laboratories will be able to join the two main components of the CDPP provided they offer a service (data dissemination or information) relating to natural plasma physics.

The archive system is currently being developed. The service is planned to be made available to the scientific community on 1/1/99.

Data Producers. Data producers are mainly either current or future experiments, or projects concerned with rehabilitating existing data. Ongoing experiments include, for example, the French experiments flown aboard Russian satellites (INTERBALL), aboard the US satellite (WIND), aboard the future European satellites (CLUSTER), or even some data from the EISCAT radars. The projects to rehabilitate existing data cover a many French experiments performed since 1975, mostly flown on European, US and Soviet satellites or probes.

II. INGEST

The CDPP has drawn up a specification for deliverable data products. The specification defines the characteristics (either mandatory or optional) that the data and metadata to be delivered to the CDPP must exhibit. It defines the rules systematically applied with respect to:

- file structure, data encoding and standardization of times and dates,
- orbit or trajectory data,
- the minimum content and format of catalogues,
- complementary information needed to use or interpret the data,
- etc.

The CDPP provides technical support in order to apply this specification to each data-producing project.

As far as future projects are concerned, the authorities empowered to make decisions on projects will make the drawing up of an obligatory data management plan. The plan must define exactly which data will be archived (physical values, raw data...), how the data will be organized, and when the data will be delivered to the CDPP.

One particular service within the CDPP is the SPID (*Service de Préparation des Informations et des Données* - Service for Preparing Information and Data), in charge of the interfaces with data-producing projects and the formatting of some metadata before its delivery to the archive system.

Submission Agreements. As far as future projects are concerned, the submission agreement shall be constituted by the project Data Management Plan, to be approved by both the project

and the CDPP. As far as existing data to be rehabilitated is concerned, the framework is less formal: there is normally no project team left and no longer a budget specific to that project. Rehabilitation is thus the responsibility of a team of engineers from CNES and those of the Principal Investigator or members of his team. The CDPP suggests priorities for the work to be completed. It also influences the choices and compromises to be made with regard to the level of data to be archived.

Delivery Session

Data delivery. Data-producing projects must normally store the data produced before delivery. They do so using the facilities offered by the STAF, a multi-mission storage service at CNES. The main function of the STAF (*Service de Transfert et d'Archivage des Fichiers* - Service for Transferring and Archiving Files) is the Long-term physical storage of information. The interface is stable and therefore the technologies and storage media can thus be replaced or changed in-house without affecting the interface. The STAF also monitors and renews the media used.

From a user project viewpoint, the STAF appears as a virtual tree structure in which files may be stored. When all the data to be delivered has been produced, the delivery process merely amounts to a change of ownership of the STAF directories in which the data is stored. There is no actual physical movement of data.

Delivery of metadata. Metadata generally takes up less space than data. A delivery disk space is set up by the CDPP and the data-producing project has the right onto write to this space. When all the data and metadata has been delivered, the SPID can begin its checking and formatting (see below). This process is valid for a complete set of data, a partial delivery or an update of previously delivered metadata.

Transformation Process

The format of experiment data is not altered during the delivery process. On the other hand, metadata delivered will be subject to a kind of packing (without changing the contents) and new metadata will be created by the SPID. To give some examples:

The archive system manages the descriptions of both data collections and objects, browse data and documentary information in the form of graphs on collections and objects. The delivery of a new collection results in the creation of a new node in the data description graph and logical links with existing collections. The creation of this information, granting a global and consistent view of all the data and metadata available, is not within the domain of the data producer.

When the Principal Investigator delivers a Microsoft Word document describing an experiment, he places the corresponding file in the delivery disk space. The SPID will then use this file to create a documentary object descriptor giving the document title, author, publishing body, language, associated keywords, stating the existence of an abstract etc.

The insertion of metadata in the archive system is mostly based on use of PVL (Parameter Value Language) and a DED (Data Entity Dictionary) which is configuration managed. One of the roles of the SPID will thus be to create this new metadata and construct the PVL structure describing it. Generally speaking, metadata appears as an extremely heterogeneous set of information objects. Using PVL means that these heterogeneous objects may be delivered in both a homogeneous and standard format.

Validation

The SPID is responsible for ensuring that the deliverable product specifications for each data set have been respected. It also performs a number of coherence checks, such as checking coherence between catalogue data and the files containing experiment data.

Once these checks have been completed, the results, together with all the metadata, are presented at a formal peer review whose purpose is to decide whether the CDPP can accept the data set and issue recommendations in this field. Once accepted, the CDPP becomes the guarantor of the data set. This review brings in scientists from outside both the CDPP and the Principal Investigator team.

Despite the various checks carried out, the scientific validity of the experiment data delivered to the CDPP remains the responsibility of the Principal Investigator or data-producing project.

Security. The delivery process for both data and metadata takes place within a dedicated environment accessible only by the data producer and the CDPP.

III. INTERNAL FORMS

Storage. The STAF multi-mission storage service (see above) takes charge of the data and metadata. This service currently uses StorageTek silos with 3490 cartridges and larger capacity Reedwood cartridges (10 Gigabytes uncompressed). The objects archived by this service are files. There are several different layers of service with regard to file retrieval time and file duplication. The STAF is in charge of all data migration involved when changing from old to new media or to a new technology medium. They do not affect the upper layers of the system.

Formats. The format of data stored must be independent of all operating systems. In practice, experiment data is usually in IEEE or ASCII code and divided up into sequential files. The application of CCSDS encoding for times and dates is compulsory for all record structure files. The syntax and semantics of each file must be described with EAST and a DED unless self-descriptive structures such as FITS or NCAR are used. As far as documentary information is concerned, no reference standard for the internal representation of documents has yet been applied.

Data Management

Data management revolves around use of a graph describing data collections and objects. For the purposes of simplification, this graph is usually known as a data graph. It is oriented and non-cyclical. The relations associating a node with its descending nodes are (from an object-oriented viewpoint) inheritance and composition relations. A data set, also known as a terminal collection, thus inherits the characteristics of all the collections above it.

Documentary information, browse data and event tables are also managed through graphs which are nonetheless distinct from the data graph. The graphs contain either explicit metadata or references to external files or documents.

IV. ACCESS

Access facilities are seen by the user through a WWW server. These facilities include aids to search for data collections and objects, means of retrieving certain metadata (such as documents and catalogues) immediately and ways of ordering data products which include special protective mechanisms for data not made public.

Finding Aids

The aids to search data of interest to the user are based on navigation within the different graphs: the experiment data collection and object graph, the browse data graph, the documentary object graph and the events table graph. These graphs are independent but a certain number of links are used to move from one to another. Navigation within the graphs is, depending on the case, through criteria such as a keyword (parameter measured, location of measurements etc.), time or other types of criteria.

The data object and collection graph grants several views of the data, and the final objects may be selected after several navigations within the graph.

The events table graph may be used to make indirect selections over time, such as selecting only data corresponding to a given instrument operating mode, or data corresponding to the periods during which a particular type of magnetospheric event was observed etc.

These aids may be used to select data which is stored either on the main archive site (at CNES) or at an associated laboratory.

Security

Without exception, metadata is visible and accessible to the general public without any prior authentication. On the other hand, data may only be ordered by a user previously authorized by the CDPP, as it normally implies the consumption of resources. The user makes his request for authorization by a form available on-line, indicating his name, e-mail address, the name of the laboratory he belongs to and the reasons for his request. Once the user has received authorization to order products, he must authenticate his request (name and password) before ordering.

Data archived by the CDPP is usually public in nature, but in the case of recent data, data ordering may be temporarily restricted to one particular user group. The system must therefore be capable of handling access rights to the service (for ordering data) independently from access rights to the data itself.

Finally, the system is designed and has a number of protective measures such that any accidental or deliberate modifications to the data stored in the Center may be avoided.

Customer Service/Support

The system can handle profiles peculiar to each user, taking into account in particular the capability of the network linking him to Internet and the laboratory to which he belongs (laboratories directly supported by CNES, laboratories involved in cooperative projects with French laboratories, other laboratories etc.).

The CDPP has a customer support team able to reply to technical questions (how to use the system, read data etc.). This team can also direct the users to the Principal Investigator or data producers.

V. DISSEMINATION

Subscriptions. In its initial version, the system only accepts orders relating to data available in the system.

Media/Network Use

The data from Plasma Physics experiments is often bulky (a data set often contains between ten and several hundred Gigabytes). It is not planned to systematically create pre-defined, widely disseminated products as is often the case for planetary data, particularly as users are often interested in a specific period of time and not the whole data set.

Products may be delivered either over a network or on a variety of media (currently CD-ROM, DAT or Exabytes). The choice between these two types of delivery depends on the capacity of the network between the user and the CDPP at any given time.

As far as network deliveries are concerned, the system proposes the HTTP protocol at the user's initiative or the FTP protocol at the CDPP's initiative, but at a time specified by the user. The latter choice is subject to certain constraints. Deliveries of data via a network offer optional data compression and grouping facilities in the form of .tar files.

Data Transformation

The data objects distributed to scientific users are not necessarily identical to the data objects stored in the system. Depending on the standards respected and tools available, a certain number of transformations of archived objects may be requested, in particular:

- Time-related retrieval which provides data corresponding to one (or more) time periods specified by the user. This kind of retrieval is only possible when times and dates have been encoded in compliance with CCSDS recommendations.
- Retrieval of fields, which permits the user to select fields of interest on the basis of an EAST data descriptor.

These transformations are known as "subsetting services". Other such transformations are planned for future versions, so as (for example) to be able to deliver data in the user's native machine format, or deliver data as physical values although it is stored as raw values.

Pricing Policy. The pricing policy has not yet been determined, but will probably include an invoice for dissemination of data on an external medium (CD-ROM, DAT, Exabytes).

Security. Whether data is public or not, data products ordered by a user are only visible and accessible by that user, whatever the mode of delivery.

ANNEX B. FEDERATION OF ARCHIVES

Users of multiple OAIS archives may have reasons to wish for some uniformity or cooperation among them. For example, Consumers of several OAIS archives may wish to have

- a master catalog to aid in locating information over several OAIS archives,
- a common schema for access and dissemination, or
- a single access site.

Producers may wish to have

- a common schema for submission to different archives, or
- a single depository for all their products.

Managers may wish to have means for

- enforcing submissions, or
- increasing the uniformity and quality of user interactions with the OAIS.

Therefore, it may be advantageous for OAIS archives to cooperate to meet these wishes. The motivation might come from the archives themselves, or it may be imposed by an authority that has some influence over them. In the former case, the archive might be motivated by the desire to keep consumers happy with their products, or to keep users happy with their quality of service, or simply by the need to compete with other archives in order to survive or grow. Situations like this can and have motivated agreements without the need for any explict federation establishing an external authorty.

In cases where explict federation is established, the external authority is represented in this Reference Model by Management. It is not the purpose of this section to discuss the detailed organizational architecture (or intrigue) of the Management interactions, since these are outside the model, but rather to describe the OAIS external interconnections that might allow the above wishes to be met.

At a rudimentary level of federation, **Figure B-1** represents a simple mutual information exchange agreement between archives. (Note: In this and the following figures, the OAIS is represented as a “five-port device” following the arrangement of Figure 4-1. In each case, a two-archive federation is shown for simplity, although the concept can be extended indefinitely.) The essential requirement for this federation is a set of mutual Submission Agreements, subscriptions, and user interface standards to allow DIPs from one archive to be ingested as SIPs by another. Therefore, it assumes that some pair-wise compatibility has between established between the archives. This does not necessarily require common access, dissemination and submission methods for all participants, although that may be expected encourage more exchange.

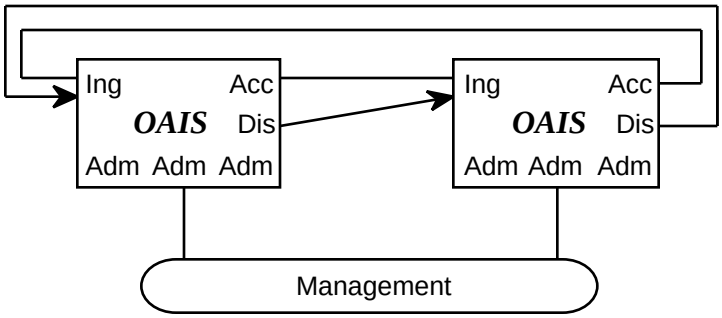


Figure B-1. A Simple OAIS Federation for Mutual Data Exchange

Figure B-2 is an example of OAIS archives that have standardized their submission and dissemination methods for the benefit of users. No special external element, other than management, is needed for this. Its disadvantage is that there is no formal mechanism for exchange of catalog information. Where does the consumer look for the desired information? Perhaps the archives agree to exchange context and catalog information, or perhaps one of the archives agrees to take on the role for both.

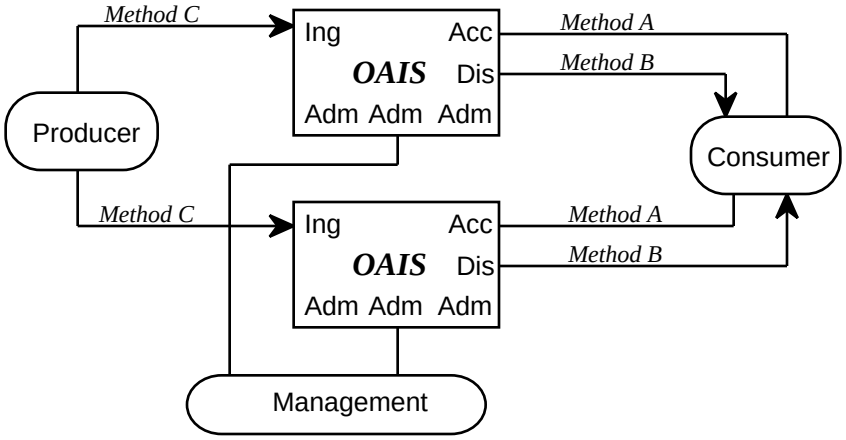


Figure B-2 An OAIS Federation with Standard Ingest and Dissemination Methods

Figure B-3 shows a way to solve the catalog problem using an entity eternal to the OAIS. Here, a pair of Producer-Consumers, each of whom maintains his own OAIS archive, have joined together to share information. The Common Catalog & Manager is the external binding element that serves as a common access point for the information in both archives. The Common Catalog may limit its activity to being a finding aid, or it may provide full combined Acces service as shown in the Figure. Optionally, it may include common Dissemination of products from either or both arcihves.

This architecture closely matches that of NASA's TIMED (Thermosphere, Mesosphere, Ionosphere Energetics and Dynamics) space mission, in which four individual archives are

operated by Principal Investigators. As Producer, each principal Investigator is responsible for archiving his own products. However, there are instances where the Producer incorporates the products of other Principal investigators into his product. In this application, the mission operator is the external entity. The Common Catalog component provides a Dissemination service which can assemble special products obtained from multiple archives, and also accept subsciptions for them. Additionally, the Manager component exploits its position to enforce the submission of data products by each Principal Investigator to his own archive, according to the agreement established for the mission.

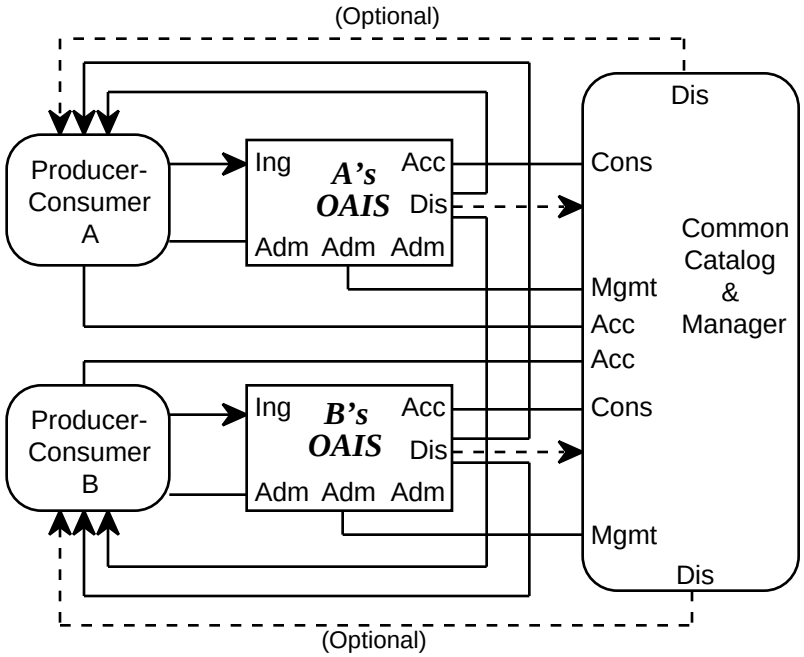


Figure B-3. An OAIS Federation Employing a Common Catalog

Figure B-4 shows a federation with an external entity on the Ingest side. As with the example of the Common Catalog, this entity permits enforcement of broad submission agreements involving several producers and archives. Here, the Common Ingest Staging entity can also take the responsibility to route submissions to the appropriate archive.

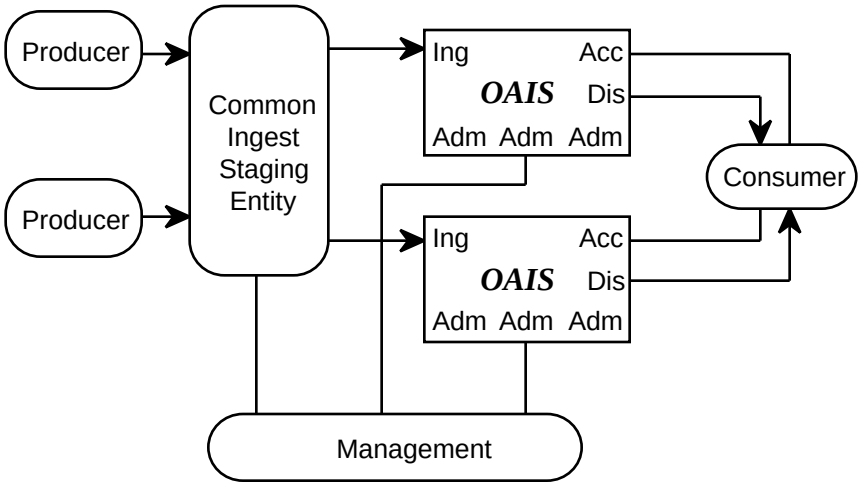


Figure B-4. An OAIS Federation with Common Ingest Staging

Of course, the arrangement of Figure B-4 is little help for the consumer who does not know where to look for information. A more elaborate arrangement than any of those shown above would include both a Common Catalog and Common Ingest Staging, to provide an external appearance similar to that of a single OAIS. Such a complete embedding of OAIS archives may be thought to be less efficient than a single distributed archive under one administration. However, it does allow for a negotiated degree of autonomy for each archive.

It should be evident from the above examples that the OAIS model is consistent with federation to accomplish specific objectives. However, it should also be considered that some of these objectives may be accomplished through voluntary action.

[Contibutor's Note: In discussing this, we may discover that implementation of federations may be made easier by adding some internal functions (perhaps in administration and access) to support them. We may even discover that some prevoiusly unrecognized means of inter-OAIS exchange among entities, such as Access-to-Access exchange, should be explicitly recognized in the model -PG]

ANNEX C. ENTITY AND FUNCTION MATRIX

Table C-1 shows the functions identified under each major entity in Section 4.1, and aligns analogous functions in rows.

Table C-1.
Sub-Functions of the OAIS Entities, Aligning Similar Functions in Rows

Entity: / Function Category	**Ingest**	**Archival Storage**	**Data Management**	**Administration**	**Access**	**Dissemination**
PLANNING	Scheduling			Planning and Scheduling		
INPUT	Staging	Transfer Receiving	Update		Provide Access Session	Receive Dissemination Request
PROCESSING	Conversion	Hierarchy Management	Report Generation		Prepare Finding Aids	Generate DIP
	Cataloging					Process Data
OUTPUT	Transfer Initiation	Provide Data	Report Production	Customer Service	Accept Dissemination Request	Delivery
QUALITY ASSURANCE	Review	Error Checking	Database Administration	Data Engineering		Monitor Dissemination Requests
		Physical Migration		Configuration Management		
SECURITY		Disaster Recovery		Physical Access Control		

ANNEX D. COMPATIBILITY WITH OTHER STANDARDS

(TBD)

ANNEX E. BRIEF GUIDE TO THE OMT

A key to object realationships in the OMT diagrams of this document is shown in Figure E-1.

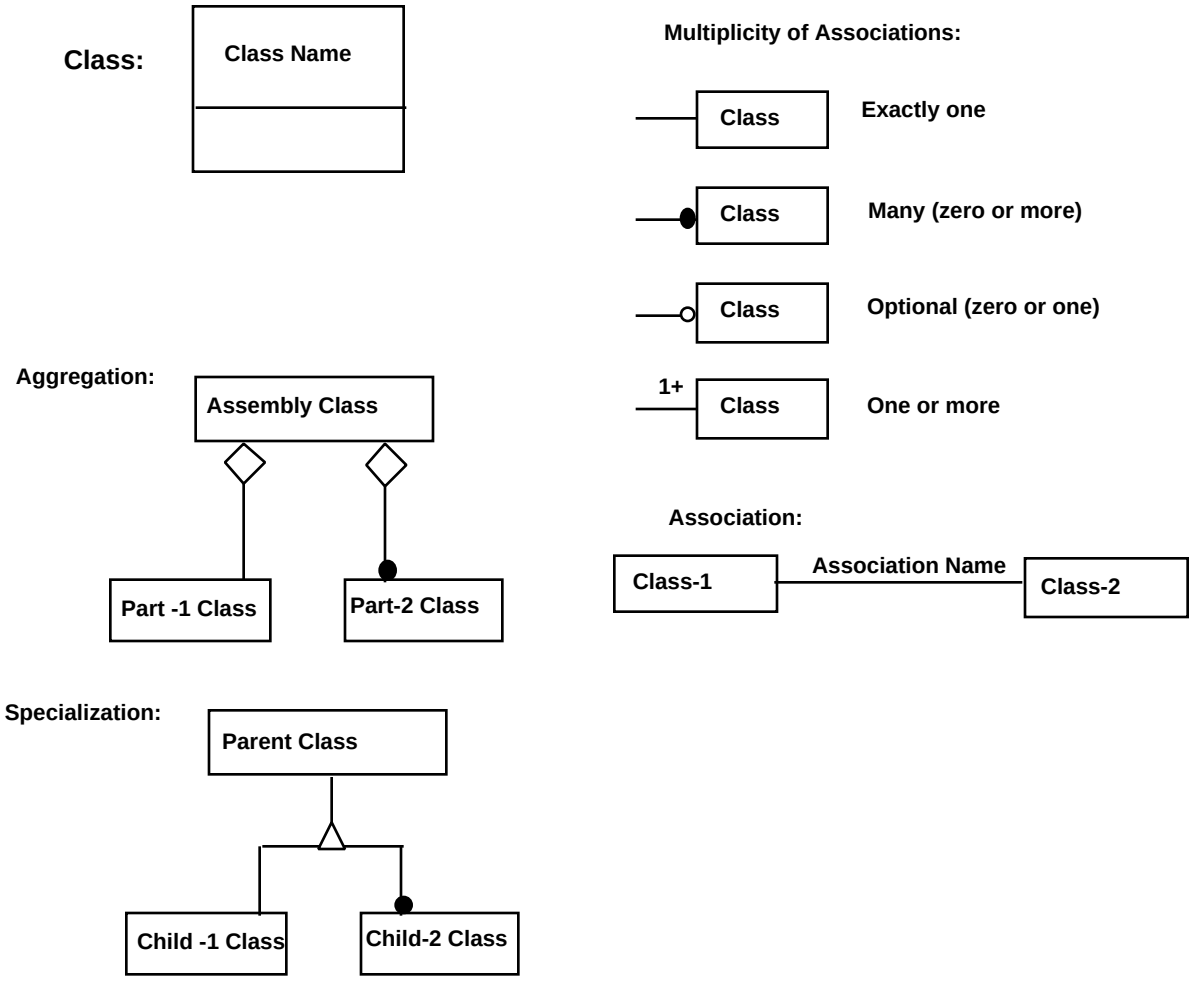


Figure E-1. Key to OMT relationships

A Class is indicated by a box, as shown, and generally there will be many possible instances of the class. Classes of objects are related to one another through Associations and there are various multiplicities that may be attached to these associations as shown. The multiplicity refers to the number of instances, or objects, of that class that are involved in the relationship.

The general association among two classes is given by a connecting line that is labeled with an information phrase indicating the nature of the association. There are two particular associations that are commonly used - aggregation and specializtion, and these have particular symbols to indicate them.

An Aggregation association is one where a class is considered to be a part of another class.

A Specialization association is one where a child class inherits properties from the parent class. A child can do what the parent can do, but generally can also do more.

In the figure, the aggregation association says that the Assembly class contains exactly one Part-1 class instance and zero or more Part-2 classe instances. The specialization association says that the Parent class properties are inherited by one instance of the Child-1 class and by zero or more instances of the Child-2 class.

*** *[End of document]* ***